PENERAPAN *COSINE SIMILARITY* PADA DETEKSI KEMIRIPAN JUDUL TUGAS AKHIR DI STMIK PALANGKARAYA

TUGAS AKHIR

Disusun Untuk Memenuhi Syarat Kelulusan Program Strata I
pada Sekolah Tinggi Manajemen Informatika dan Komputer
(STMIK) Palangkaraya

OLEH

IBNU ROFIK
C2255201109
PROGRAM STUDI TEKNIK INFORMATIKA

SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN KOMPUTER
(STMIK) PALANGKARAYA
2026

PENERAPAN *COSINE SIMILARITY* PADA DETEKSI KEMIRIPAN JUDUL TUGAS AKHIR DI STMIK PALANGKARAYA

TUGAS AKHIR

Disusun Untuk Memenuhi Syarat Kelulusan Program Strata I
pada Sekolah Tinggi Manajemen Informatika dan Komputer
(STMIK) Palangkaraya

OLEH

IBNU ROFIK
C2255201109
PROGRAM STUDI TEKNIK INFORMATIKA

SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN KOMPUTER
(STMIK) PALANGKARAYA
2026

PERSETUJUAN

PENERAPAN *COSINE SIMILARITY* PADA DETEKSI KEMIRIPAN JUDUL TUGAS AKHIR DI STMIK PALANGKARAYA

Tugas Akhir Ini Telah Disetujui dan Disahkan

Pembimbing I,	Pembimbing II.
Dr. Maura Widyaningsih, S.Kom,M.Cs	Ir. Abdul Hadi, S.T., M.Kom.
NIK. 112233445566778	NIK. 197502232009002

Mengetahui	Mengesahkan
Ketua STMIK Palangkaraya,	Ketua Prodi Teknik Informatika,
Ir. Hj. Siti Maryamah, M.M	Ir. Abdul Hadi, S.T., M.Kom
NIK. 196902021995004	NIK. 197502232009002

PENGESAHAN

PENERAPAN *COSINE SIMILARITY* PADA DETEKSI KEMIRIPAN JUDUL TUGAS AKHIR DI STMIK PALANGKARAYA

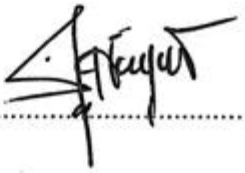
Tugas Akhir ini telah Diujikan, Dinilai, dan Disahkan
Oleh Tim Penguji pada Tanggal 26 Mei 2026

Tim Penguji Tugas Akhir :

1. Herkules, S.Kom., M.Cs.
   Ketua ..... 

2. Sherly Jayanti, S.T., M.Cs
   Sekretaris ..... 

3. Sulistyowati. S.Kom., M.Cs
   Anggota ..... 

4. Dr. Maura. Widyaningsih, S.Kom., M.Cs
   Anggota ..... 

5. Ir. Abdul Hadi, S.T., M.Kom
   Anggota ..... 

PERNYATAAN KEASLIAN

Yang bertanda tangan di bawah ini:

Nama Mahasiswa	:	Ibnu Rofik
N I M	:	C2255201109

menyatakan bahwa Tugas Akhir dengan judul :

PENERAPAN *COSINE SIMILARITY* PADA DETEKSI KEMIRIPAN JUDUL TUGAS AKHIR DI STMIK PALANGKARAYA

adalah hasil karya saya dan bukan merupakan duplikasi sebagian atau seluruhnya dari karya orang lain, kecuali bagian yang sumber informasi dicantumkan.

Pernyataan ini dibuat dengan sebenar-benarnya secara sadar dan bertanggungjawab dan saya bersedia menerima sanksi pembatalan Tugas Akhir apabila terbukti melakukan duplikasi terhadap Tugas Akhir atau karya ilmiah lain yang sudah ada.

Palangka Raya, 26 Mei 2026

Yang Membuat Pernyataan,



IBNU ROFIK

MOTTO DAN PERSEMBAHAN

"*if only we started younger, then we would have been so much better, if we know back then, what we did now? things would change.*"

Dengan penuh rasa syukur kepada Allah SWT. Tugas akhir ini saya persembahkan kepada:

1. Kedua orang tua tercinta yang senantiasa memberikan doa, dukungan, kasih sayang, serta semangat sehingga penulis dapat menyelesaikan tugas akhir ini.
2. Dr. Maura Widyaningsih dan Ir. Abdul Hadi selaku dosen pembimbing yang telah memberikan arahan, bimbingan, dan motivasi kepada penulis selama proses penyusunan tugas akhir.
3. Seluruh keluarga, sahabat, dan teman-teman yang telah memberikan dukungan, bantuan, serta semangat kepada penulis dalam menyelesaikan tugas akhir ini.

INTISARI

Ibnu Rofik, C2255201109, 2026. Penerapan *Cosine Similarity* Pada Deteksi Kemiripan Judul Tugas Akhir Di STMIK Palangkaraya, Pembimbing I Dr. Maura Widyaningsih, S.Kom,M.Cs, M.Kom., Pembimbing II Ir. Abdul Hadi, S.T., M.Kom.

Proses verifikasi pengajuan judul Tugas Akhir (TA) di STMIK Palangkaraya saat ini masih dilakukan secara manual. Seiring meningkatnya volume repositori setiap tahunnya, metode ini kurang efektif dan rentan terhadap duplikasi topik. Penelitian ini bertujuan membangun sistem deteksi kemiripan judul TA berbasis *web* yang terintegrasi secara otomatis guna mempermudah mahasiswa dan dosen memverifikasi orisinalitas judul.

Sistem dikembangkan menggunakan model *Waterfall*. Data judul ditarik otomatis dari website perpustakaan menggunakan teknik *web scraping* berbasis *Python*, lalu disimpan dalam database *Google Spreadsheets*. Data teks melalui tahapan *preprocessing* meliputi *case folding*, *tokenizing, stopword removal*, dan *stemming* berbantuan Sastrawi.js. Algoritma *TF-IDF* digunakan sebagai metode pembobotan kata dan *Cosine Similarity* untuk menghitung persentase derajat kemiripan antar-judul.

Hasil pengujian *black box* menunjukkan seluruh fungsi sistem berjalan dengan baik. Berdasarkan uji validasi algoritma terhadap 30 pasangan data judul uji yang dibandingkan dengan penilaian dosen pakar (*ground truth*), sistem memperoleh akurasi sebesar 73,3% menggunakan *threshold* awal 71% yang mengacu pada penelitian sebelumnya. *Threshold* tersebut kemudian dikalibrasi menjadi 60% berdasarkan temuan bahwa banyak pasangan judul yang dinilai Tinggi oleh pakar berada pada rentang kemiripan 60%-71%, sehingga *threshold* 71% kurang optimal untuk konteks kemiripan judul. Setelah kalibrasi, akurasi sistem meningkat menjadi 83,3%.

Kata kunci: *Cosine Similarity*, *TF-IDF*, Tugas Akhir, *Web Scraping*, Preprocessing Teks

ABSTRACT

Ibnu Rofik, C2255201109, 2026. Application of Cosine Similarity for Detecting Similarity of Final Project Titles at STMIK Palangkaraya, Advisor I Dr. Maura Widyaningsih, S.Kom., M.Cs., Advisor II Ir. Abdul Hadi, S.T., M.Kom.

The verification process for submitting Final Project (TA) titles at STMIK Palangkaraya is currently conducted manually. As the repository volume increases annually, this method becomes inefficient and prone to topic duplication. This research aims to build a web-based final project title similarity detection system that integrates automatically to assist students and lecturers in verifying title originality.

The system was developed using the Waterfall model. Title data was collected automatically from the library website using Python-based web scraping techniques and stored in a Google Spreadsheets database. The text data went through preprocessing stages, including case folding, tokenizing, stopword removal, and stemming using Sastrawi.js. The TF-IDF algorithm was used for word weighting and Cosine Similarity to calculate the similarity percentage between titles.

Black box testing results indicate that all system functions performed well. Based on algorithm validation testing on 30 pairs of test title data compared with expert lecturer assessments (ground truth), the system achieved an accuracy of 73.3% using an initial threshold of 71% referring to previous research. The threshold was then calibrated to 60% based on the finding that many title pairs rated High by experts were in the 60%-71% similarity range, making the 71% threshold suboptimal for the title similarity context. After calibration, the system accuracy increased to 83.3%.

Keywords: Cosine Similarity, TF-IDF, Final Project, Web Scraping, Text Preprocessing

PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR

Sebagai sivitas akademik Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) Palangkaraya, saya yang bertanda tangan dibawah ini:

Nama : Ibnu Rofik

NIM : C2255201109

Program Studi : Teknik Informatika

Jenis Karya : Laporan Tugas Akhir

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) Palangkaraya, Hak Bebas Royalti Non Eksklusif (*Non-exclusive Royalty-Free Right*) atas karya ilmiah saya yang berjudul:

PENERAPAN COSINE SIMILARITY PADA DETEKSI KEMIRIPAN JUDUL TUGAS AKHIR DI STMIK PALANGKARAYA

Dengan Hak Bebas Royalti Non Eksklusif ini Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) Palangkaraya berhak menyimpan, mengalih media/formatkan, mengelola dalam bentuk pangkalan data (database), merawat, dan mempublikasikan tugas akhir saya selama tetap mencantumkan nama saya penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di : Palangka Raya
Pada tanggal : 26 Mei 2026

Yang menyatakan



Ibnu Rofik

KATA PENGANTAR

Puji dan syukur penulis panjatkan ke hadirat Tuhan Yang Maha Esa karena atas rahmat dan karunia-Nya penulis dapat menyelesaikan Tugas Akhir ini dengan baik. Penyusunan Tugas Akhir ini merupakan salah satu syarat untuk menyelesaikan pendidikan strata satu dan memperoleh gelar sarjana di STMIK Palangkaraya.

Penulis mengucapkan terima kasih kepada berbagai pihak yang telah memberikan bantuan, dukungan, serta bimbingan baik secara langsung maupun tidak langsung selama proses penyusunan Tugas Akhir ini, terutama kepada:

1. Dr. Maura Widyaningsih, S.Kom, M.Cs selaku dosen pembimbing I yang telah meluangkan banyak waktu, memberikan arahan dan bimbingan selama penyusunan Tugas Akhir ini.
2. Ir. Abdul Hadi, S.T., M.Kom. selaku dosen pembimbing II yang telah memberikan arahan dan bimbingan selama penyusunan Tugas Akhir ini.
3. Serta orang tua dan teman-teman yang mendukung penulis untuk bisa menyelesaikan tugas akhir ini.

Penulis Mengharapkan Kritik dan saran yang bersifat membangun untuk dapat menyempurnakan Tugas Akhir ini. Semoga Tugas Akhir ini dapat bermanfaat bagi kita semua terutama untuk menambah ilmu pengetahuan.

Palangka Raya, 26 Mei 2026

Penulis

DAFTAR ISI

PENGESAHAN ..... **iv**
PERNYATAAN KEASLIAN ..... **iv**
MOTTO DAN PERSEMBAHAN ..... **vi**
INTISARI ..... **vii**
ABSTRACT ..... **viii**
KATA PENGANTAR ..... **x**
DAFTAR ISI ..... **xi**
DAFTAR TABEL ..... **xiii**
DAFTAR GAMBAR ..... **xiv**
DAFTAR LAMPIRAN ..... **xv**
BAB I PENDAHULUAN **1**
1.1 Latar Belakang Masalah 1
1.2 Perumusan Masalah 3
1.3 Batasan Masalah 4
1.4 Tujuan dan Manfaat 5
1.5 Sistematika Penulisan 5
BAB II TINJAUAN PUSTAKA **7**
2.1 Kajian Teori 7
2.1.1 Tugas Akhir 7
2.1.2 Tugas Akhir 7
2.1.3 *Text Mining* 8
2.1.4 *Web Scraping* 12
2.1.5 Term Frequency–Inverse Document Frequency (TF-IDF) 12
2.1.6 *Cosine Similarity* 14
2.1.7 *Waterfall* 15
2.1.8 Perangkat Lunak dan Alat yang digunakan 17
2.2 Penelitian Yang Relevan 21
BAB III METODOLOGI PENELITIAN **25**
3.1 Tinjauan Umum 25
3.2 Jenis Penelitian 26
3.3 Penelitian Yang Relevan 26
3.4 Instrumen penelitian 29
3.5 Analisis Kebutuhan 29
3.5.1 Analisis Data 30
3.5.2 Analisis Proses 38
3.5.3 Analisis Kelemahan 41
3.6 Desain 41
BAB IV PEMBAHASAN **51**
4.1 Hasil 51
4.1.1 Implementasi Sistem 51
4.1.2 Implementasi Sistem 58
4.2 Pengujian 64
4.2.1 Uji *Blackbox* 64
4.2.2 Uji Validasi Algoritma 65

4.3 Pembahasan 74
BAB V KESIMPULAN DAN SARAN 77
5.1 Kesimpulan 77
5.2 Saran 78
DAFTAR PUSTAKA

DAFTAR TABEL

Tabel 2.1 Contoh Proses *Case Folding* 8
Tabel 2.2 Contoh Proses *Tokenizing* 9
Tabel 2.3 Contoh Proses *Stopword Removal* 10
Tabel 2.4 Contoh Proses *Stemming* 11
Tabel 2.5 Penelitian Relevan 21
Tabel 3.1 Perangkat Keras 30
Tabel 3.2 Perangkat Lunak 30
Tabel 3.3 Tingkat Kemiripan Judul 31
Tabel 3.4 Hasil *Case Folding* pada Judul Tugas Akhir 32
Tabel 3.5 Hasil *Tokenizing* pada Judul Tugas Akhir 32
Tabel 3.6 Hasil *Stopword Removal* pada Judul Tugas Akhir 33
Tabel 3.7 Hasil *Stemming* pada Judul Tugas Akhir 34
Tabel 3.8 *Frekuensi Kemunculan Term* (TF) pada Setiap Dokumen 34
Tabel 3.9 Perhitungan Nilai IDF untuk Setiap *Term* 35
Tabel 3.10 Nilai Bobot TF-IDF Setiap *Term* pada Masing-Masing Dokumen 36
Tabel 3.11 Hasil Akhir *Cosine Similarity* 38
Tabel 3.12 *Database Spreadsheets* 50
Tabel 4.1 Hasil Akhir Token Manual dan Sistem 60
Tabel 4.2 Hasil *Cosine Similarity* pada Sistem 63
Tabel 4.3 Perbandingan *Cosine Similarity* Sistem dan Manual 64
Tabel 4.4 Uji *blackbox* pada sistem *website* 64
Tabel 4.5 Uji *Blackbox* pada aplikasi scraping 65
Tabel 4.6 Hasil Penilaian Kemiripan oleh pakar 66
Tabel 4.7 *Confussion Matrix* 71
Tabel 4.8 Hasil Kinerja Sistem 72
Tabel 4.9 *Confusion Matrix* Setelah Penyesuaian 72
Tabel 4.10 Hasil Kinerja Sistem Setelah Penyesuaian 73
Tabel 4.11 Perbandingan Kinerja Sistem sebelum dan Sesudah Penyesuaian 73

DAFTAR GAMBAR

Gambar 2.1 Metode *Waterfall* 16
Gambar 3.1 Flowchart Alur Kerja *Website* 39
Gambar 3.2 *Flowchart* Alur Kerja *Scraping* 40
Gambar 3.3 *Usecase Diagram* proses 41
Gambar 3.4 *Activity Diagram* Alur Sistem 42
Gambar 3.5 *Sequence Diagram* Admin 43
Gambar 3.6 *Sequence Diagram* Pengguna 44
Gambar 3.7 Desain Halaman Utama 45
Gambar 3.8 Desain Halaman Hasil 46
Gambar 3.9 Desain Halaman Detail 47
Gambar 3.10 Desain Halaman Daftar Judul 49
Gambar 4.1 Proses *Scraping* 51
Gambar 4.2 Hasil *Scraping* Data Judul 52
Gambar 4.3 *Upload* File xlsx ke *Spreadsheets* 53
Gambar 4.4 *Upload* Data Judul ke Publik 54
Gambar 4.5 Lokasi GID 54
Gambar 4.6 Halaman Utama 55
Gambar 4.7 Halaman Hasil 56
Gambar 4.8 Halaman Detail 57
Gambar 4.9 Halaman Referensi 57
Gambar 4.10 *Stopwords Database* 59
Gambar 4.11 Kamus tambahan *Stemming* 60

DAFTAR LAMPIRAN

Lampiran 1. Surat Tugas Pembimbing Tugas Akhir
Lampiran 2. Surat Tugas Penguji Sidang Tugas Akhir
Lampiran 3. Kartu Kegiatan Konsultasi Tugas Akhir
Lampiran 4. Berita Acara Ujian Tugas Akhir
Lampiran 5. Arsip Data Penelitian Dan Repositori

BAB I
PENDAHULUAN

1.1 Latar Belakang Masalah

Tugas akhir merupakan salah satu syarat kelulusan bagi mahasiswa STMIK Palangkaraya. Dalam proses pengajuan judul, sering ditemui permasalahan berupa kemiripan atau kesamaan judul antara mahasiswa dengan penelitian terdahulu. Kemiripan judul tidak hanya mengurangi orisinalitas karya ilmiah, tetapi juga berpotensi menimbulkan duplikasi topik serta tumpang tindih substansi penelitian. Hal ini sejalan dengan pendapat (Purwaningrum et al., 2025) yang menegaskan bahwa tingginya tingkat kemiripan judul tugas akhir dapat menimbulkan berbagai permasalahan, seperti indikasi plagiarisme akademik dan redundansi penelitian yang merugikan dalam konteks pengembangan ilmu pengetahuan.

Berdasarkan data dari *website* Perpustakaan STMIK Palangkaraya, tercatat sebanyak 2316 judul tugas akhir yang tersimpan dalam repositori hingga tahun 2025. Jumlah tersebut menunjukkan bahwa volume arsip penelitian yang dikelola cukup besar dan terus bertambah seiring dengan meningkatnya jumlah mahasiswa. Peningkatan jumlah data tersebut secara langsung berdampak pada bertambahnya beban kerja dosen pembimbing dan koordinator tugas akhir dalam proses verifikasi judul. Pada praktiknya, proses pemeriksaan kemiripan judul masih dilakukan secara manual dengan cara mencocokkan judul baru terhadap daftar judul lama satu per satu, sehingga memerlukan waktu yang lama dan berpotensi menimbulkan ketidaktepatan dalam pengambilan keputusan.

Meskipun *website* resmi perpustakaan kampus telah menyediakan arsip Tugas Akhir secara online, sistem tersebut belum dilengkapi dengan fitur pencarian kemiripan judul secara otomatis. Akses data hanya berupa tampilan informasi judul dan dokumen, tanpa kemampuan analisis berbasis teks yang dapat mendukung proses verifikasi. Penelitian oleh (Salman, 2025), menegaskan bahwa digitalisasi arsip tanpa disertai mekanisme pencarian dan analisis otomatis masih belum cukup untuk meningkatkan efisiensi kerja, karena proses peninjauan tetap harus dilakukan secara manual oleh dosen atau pengelola repositori. Ketiadaan fitur analisis otomatis inilah yang membuat proses pengecekan judul di STMIK Palangkaraya rentan mengalami keterlambatan dan ketidakefektifan.

Selain itu, *website* Perpustakaan STMIK Palangkaraya tidak menyediakan *API* resmi atau *endpoint* integrasi untuk mengakses data secara otomatis, sehingga pembaruan data harus dilakukan dengan cara *web scraping* secara periodik. Kondisi ini menimbulkan kebutuhan akan sistem yang mampu mengotomatisasi pengambilan data, membersihkan data (*preprocessing*), serta melakukan perhitungan kemiripan judul menggunakan algoritma khusus. Dengan demikian, sistem pendeteksi kemiripan judul tidak bergantung pada input manual dan mampu memperbarui *database* secara berkala sesuai dengan perkembangan arsip penelitian.

Untuk mengatasi permasalahan tersebut, diperlukan sistem yang dapat mendeteksi kemiripan judul tugas akhir secara otomatis menggunakan algoritma *Term Frequency–Inverse Document Frequency* (TF-IDF) dan *Cosine Similarity*. Menurut (Nasrullah, 2024) algoritma *Cosine Similarity* memiliki keunggulan dalam

mengukur kedekatan antar teks karena mampu menghitung derajat kemiripan berdasarkan representasi vektor, sehingga tepat digunakan dalam proses pencarian dan deteksi kemiripan dokumen. Sementara itu, TF-IDF efektif dalam memberikan bobot pada kata-kata penting sehingga sistem dapat mengenali istilah dominan dalam judul penelitian. Kombinasi kedua algoritma ini terbukti banyak digunakan dalam sistem rekomendasi, pencarian dokumen, dan deteksi plagiarisme.

Dengan adanya sistem deteksi kemiripan judul yang terintegrasi dengan repositori digital kampus, proses verifikasi judul dapat dilakukan dengan lebih cepat, akurat, dan efisien. Sistem ini tidak hanya membantu dosen dalam mengurangi beban pemeriksaan manual, tetapi juga memudahkan mahasiswa untuk mengetahui tingkat kesesuaian atau potensi kemiripan judul yang mereka ajukan. Selain itu, penggunaan teknologi seperti *web scraping*, preprocessing teks otomatis, serta penyimpanan data yang terpusat pada platform digital dengan *Google Sheets* yang memungkinkan sistem tetap ringan, mudah dipelihara, dan dapat diperbarui tanpa proses *input* data satu per satu. Oleh karena itu, pengembangan sistem pendeteksi kemiripan judul berbasis *web* menjadi solusi yang relevan dan diperlukan dalam mendukung manajemen tugas akhir di STMIK Palangkaraya.

1.2 Perumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, maka rumusan masalah dalam penelitian ini adalah bagaimana merancang dan membangun sistem berbasis *web* untuk mendeteksi kemiripan judul tugas akhir, serta bagaimana mengimplementasikan algoritma TF-IDF dan *Cosine Similarity* dalam proses pendeteksian tingkat kemiripan judul tugas akhir pada STMIK Palangkaraya.

Selain itu, penelitian ini juga membahas pembangunan aplikasi *web scraping* untuk pengambilan dan pembaruan data judul tugas akhir dari *website* Perpustakaan STMIK Palangkaraya.

1.3 Batasan Masalah

Agar penelitian ini lebih terarah dan fokus, maka ditetapkan batasan masalah sebagai berikut:

1. Sistem yang dibangun berbasis *web* dan hanya digunakan untuk mendeteksi kemiripan judul tugas akhir serta menyimpan metadata tugas akhir.
2. Data judul tugas akhir yang digunakan adalah data judul yang telah terdaftar dan tersimpan di *website* Perpustakaan STMIK Palangkaraya.
3. Algoritma yang digunakan untuk mendeteksi kemiripan judul adalah *Term Frequency-Inverse Document Frequency* (TF-IDF) dan *Cosine Similarity*.
4. Sistem mendeteksi kemiripan berdasarkan teks judul tugas akhir, tidak menganalisis isi lengkap dari dokumen tugas akhir.
5. Pembaruan data dilakukan melalui proses *web scraping* yang dijalankan oleh admin perpustakaan untuk mengambil data terbaru dari *website* Perpustakaan STMIK Palangkaraya.

1.4 Tujuan dan Manfaat

a. Tujuan

Tujuan dari penelitian ini adalah merancang dan membangun sebuah sistem berbasis *website* untuk mendeteksi kemiripan tugas akhir di STMIK Palangkaraya dengan metode TF-IDF dan *Coosine Similarity*.

b. Manfaat

Penelitian ini diharapkan dapat memberikan manfaat sebagai berikut:

1) Bagi Penulis

Penelitian ini dapat menambah pengalaman dan portofolio penulis dalam melakukan penelitian, menganalisis permasalahan, serta mengembangkan aplikasi berbasis *web* dengan menerapkan metode TF-IDF dan *Cosine Similarity* untuk mendeteksi kemiripan judul tugas akhir.

2) Bagi Pengguna

penelitian ini diharapkan dapat membantu mahasiswa dan dosen dalam melakukan pengecekan kemiripan judul tugas akhir secara lebih cepat dan efisien, sehingga dapat meminimalkan terjadinya pengajuan judul yang memiliki tingkat kemiripan tinggi dengan judul yang telah ada sebelumnya.

1.5 Sistematika Penulisan

Adapun sistematika penulisan dalam penulisan tugas akhir ini terdiri dari bebrapa bab dan masing-masing bab membahas dan menguraikan pokok permasalahan yang berbeda, sebagai gambaran disini penulis menyertakan garis-garis besarnya yaitu :

BAB I PENDAHULUAN

Bab ini berisikan latar belakang masalah, rumusan masalah, Batasan masalah, tujuan dan manfaat, serta sistematika penulisan.

BAB II LANDASAN TEORI

Bab ini berisikan tentang tinjauan pustaka yang diambil dari penelitian yang relevan beserta susunan kajian teori yang disesuaikan dengan tema Tugas Akhir.

BAB III METODOLOGI PENELITIAN

Bab ini berisikan tentang tahapan yang dilakukan peneliti dalam mengumpulkan informasi atau data yang dibutuhkan.

BAB IV HASIL DAN PEMBAHASAN

Bab ini berisikan tentang hasil dari penelitian dan pembahasan dari hasil yang didapat.

BAB V KESIMPULAN DAN SARAN

Bab ini berisikan tentang kesimpulan dan saran yang diberikan untuk pengembangan selanjutnya

BAB II
TINJAUAN PUSTAKA

2.1 Kajian Teori

2.1.1 Tugas Akhir

Berdasarkan Pedoman Akademik STMIK Palangkaraya (2022), tugas akhir merupakan mata kuliah wajib yang harus ditempuh mahasiswa sebagai syarat kelulusan, dengan bentuk dan bobot SKS yang disesuaikan oleh masing-masing program studi. Bagi mahasiswa Diploma, tugas akhir berupa proyek atau paper, sedangkan bagi mahasiswa Program Sarjana berbentuk skripsi.

Menurut Buku Pedoman Tugas Akhir Program Studi Teknik Informatika (2025), pelaksanaan Tugas Akhir memiliki tujuan untuk melatih mahasiswa dalam melakukan kegiatan penelitian secara menyeluruh, yang meliputi pemilihan topik, penyusunan rencana penelitian, pelaksanaan penelitian, evaluasi hasil, serta penulisan dan pertanggungjawaban laporan penelitian secara akademik.

2.1.2 Tugas Akhir

Website adalah suatu aplikasi yang diakses menggunakan *web browser* melalui suatu jaringan seperti internet atau intranet. *Website* dapat diartikan suatu kumpulan-kumpulan halaman yang menampilkan berbagai macam informasi teks, data, gambar, animasi, suara, video maupun gabungan dari semuanya, baik statis maupun dinamis, yang dimana membentuk satu rangkaian bangunan yang saling berkaitan dan masing-masing dihubungkan dengan jaringan halaman atau *hyperlink* (Amanah Amnun Zulfa, Tatang Ibrahim, 2025)

Dalam penelitian ini, *website* berperan sebagai *platform* utama untuk menampilkan hasil pengecekan judul serta menyediakan antarmuka bagi pengguna, untuk mahasiswa, dalam melakukan input judul tugas akhir.

2.1.3 *Text Mining*

a. Menurut Halim (Halim, 2024), *Text mining* merupakan suatu proses yang melibatkan penerapan berbagai metode, alat, dan teknik analisis untuk mengekstraksi serta menemukan pengetahuan dari kumpulan dokumen teks dalam jumlah besar. Pengetahuan yang diperoleh melalui proses ini dihasilkan dari pengolahan dan ekstraksi informasi pada sumber data, sehingga menghasilkan kata-kata informatif yang mencerminkan pola serta kecenderungan statistik dalam teks tersebut.

b. Sebelum proses perhitungan dapat dilakukan, data teks perlu melalui tahap preprocessing untuk memastikan bahwa teks berada dalam bentuk yang bersih, terstruktur, dan siap diproses oleh algoritma. Tahapan ini penting agar hasil analisis lebih akurat terhadap judul. Adapun tahapan umum *preprocessing* dalam *text mining* meliputi:

    a. Case Folding

c. Tahap ini bertujuan menyeragamkan format teks dengan mengubah seluruh huruf menjadi huruf kecil (*lowercase*) agar sistem tidak membedakan huruf besar dan kecil. Contoh proses *Case Folding* ditunjukkan pada Tabel 2.1.

Tabel 2.1 Contoh Proses *Case Folding*

Teks Asli	Hasil *Case Folding*
Markas Besar Kepolisian Republik Indonesia melalui Divisi Teknologi	markas besar kepolisian republik indonesia melalui divisi teknologi

Teks Asli	Hasil *Case Folding*
Informasi Div TI Polri meluncurkan aplikasi berbasis Android dan iOS yang bernama PolisiKu	informasi div ti polri meluncurkan aplikasi berbasis android dan ios yang bernama polisiku
Badan Kepegawaian Negara BKN mengklarifikasi atas beredarnya surat pengangkatan Calon Pegawai Negeri Sipil CPNS di Palembang yang dilansir dari dan merupakan surat ilegal atau palsu	badan kepegawaian negara bkn mengklarifikasi atas beredarnya surat pengangkatan calon pegawai negeri sipil cpns di palembang yang dilansir dari dan merupakan surat ilegal atau palsu
Saya mendukung Prabowo Subianto karena Indonesia saat ini butuh pemimpin yang tegas dan ikhlas untuk mengangkat rakyat dari keterpurukan	saya mendukung prabowo subianto karena indonesia saat ini butuh pemimpin yang tegas dan ikhlas untuk mengangkat rakyat dari keterpurukan

(Sumber: Kurniawan and Mustikasari, 2021)

b. *Tokenizing*

Tokenizing merupakan proses pemecahan teks atau kalimat menjadi bagian-bagian yang lebih kecil yang disebut token. Proses ini bertujuan untuk memudahkan analisis setiap kata secara individual dengan menyusunnya ke dalam bentuk *array*. Contoh hasil proses *tokenizing* ditunjukkan pada Tabel 2.2.

Tabel 2.2 Contoh Proses *Tokenizing*

Teks Input	Hasil *Tokenizing*
markas besar kepolisian republik indonesia melalui divisi teknologi informasi div ti polri meluncurkan aplikasi berbasis android dan ios yang bernama polisiku	["markas", "besar", "kepolisian", "republik", "indonesia", "melalui", "divisi", "teknologi", "informasi", "div", "ti", "polri", "meluncurkan", "aplikasi", "berbasis", "android", "dan", "ios","yang", "bernama", "polisiku"]
badan kepegawaian negara bkn mengklarifikasi atas beredarnya surat pengangkatan calon pegawai negeri sipil cpns di palembang yang dilansir dari dan merupakan surat ilegal atau palsu	["badan", "kepegawaian", "negara", "bkn", "mengklarifikasi", "atas", "beredarnya", "surat", "pengangkatan", "calon", "pegawai", "negeri", "sipil", "cpns", "di", "palembang", "yang",

Teks Input	Hasil *Tokenizing*
	"dilansir", "dari", "dan", "merupakan", "surat", "ilegal", "atau", "palsu"]
saya mendukung prabowo subianto karena indonesia saat ini butuh pemimpin yang tegas dan ikhlas untuk mengangkat rakyat dari keterpurukan	["saya", "mendukung", "prabowo", "subianto", "karena", "indonesia", "saat", "ini", "butuh", "pemimpin", "yang", "tegas", "dan", "ikhlas", "untuk", "mengangkat", "rakyat", "dari", "keterpurukan"]

(Sumber: Kurniawan and Mustikasari, 2021)

c. Stopwords Removal

Tahap ini dilakukan untuk menghapus kata-kata yang dianggap tidak memiliki makna penting atau tidak berpengaruh terhadap proses analisis. Tujuannya adalah agar hanya kata-kata yang bersifat informatif yang diproses lebih lanjut. Pada penelitian ini, proses *stopword removal* menggunakan daftar *stopword* bahasa Indonesia yang diperoleh dari dataset publik Kaggle. Contoh hasil proses *stopword removal* ditunjukkan pada Tabel 2.3.

Tabel 2.3 Contoh Proses Stopword Removal

Kata Stopword Removal	Kata Setelah *Stopword Removal*
markas besar kepolisian republik indonesia melalui divisi teknologi informasi div ti polri meluncurkan aplikasi berbasis android dan ios yang bernama polisiku	["markas", "kepolisian", "republik", "indonesia", "divisi", "teknologi", "informasi", "div", "ti", "polri", "meluncurkan", "aplikasi", "berbasis", "android", "ios", "bernama", "polisiku"]
badan kepegawaian negara bkn mengklarifikasi atas beredarnya surat pengangkatan calon pegawai negeri sipil cpns di palembang yang dilansir dari dan merupakan surat ilegal atau palsu	"badan", "kepegawaian", "negara", "bkn", "mengklarifikasi", "beredarnya", "surat", "pengangkatan", "calon", "pegawai", "negeri", "sipil", "cpns", "palembang", "dilansir", "surat", "ilegal", "palsu"]
saya mendukung prabowo subianto karena indonesia saat ini butuh pemimpin yang tegas dan ikhlas	["mendukung", "prabowo", "subianto", "indonesia", "butuh", "pemimpin", "ikhlas",

Kata Stopword Removal	Kata Setelah *Stopword Removal*
untuk mengangkat rakyat dari keterpurukan	"mengangkat", "rakyat", "keterpurukan"]

(Sumber: Kurniawan and Mustikasari, 2021)

d. *Stemming*

Stemming merupakan proses mengubah setiap kata ke bentuk dasarnya. Tahap ini bertujuan untuk mengurangi variasi bentuk kata yang memiliki makna sama sehingga meminimalkan duplikasi dalam analisis teks. Pada penelitian ini, proses stemming dilakukan menggunakan *library* Sastrawi.js, yaitu implementasi *stemming* Bahasa Indonesia berbasis *JavaScript*. Contoh hasil proses *stemming* ditunjukkan pada Tabel 2.4

Tabel 2.4 Contoh Proses *Stemming*

Kata Sebelum Stemming	Kata Setelah *Stemming*
['presiden', 'jokowi', 'perintahkan', 'menterinya', 'setop', 'wacana', 'presiden', 'periode', 'diamdiam', 'bocoran', 'informasi', 'dukungan', 'presiden', 'periode', 'digas', 'momentum', 'peringatan', 'lahir', 'pancasilan']	['presiden', 'jokowi', 'perintah', 'menteri', 'setop', 'wacana', 'presiden', 'periode', 'diamdiam', 'bocor', 'informasi', 'dukung', 'presiden', 'periode', 'gas', 'momentum', 'ingat', 'lahir', 'pancasilan']
['komitmen', 'presiden', 'bersuara', 'penundaan', 'pemilu', 'jabatan', 'periode', 'bikin', 'gaduh']	['komitmen', 'presiden', 'suara', 'tunda', 'milu', 'jabat', 'periode', 'bikin', 'gaduh']
['copot', 'mentri', 'ri', 'mendukung', 'presiden', 'periode']	['copot', 'tri', 'ri', 'dukung', 'presiden', 'periode']
Analisis Sentimen Pengguna Media Sosial Twitter	analisis sentimen guna media sosial twitter
Deteksi Kemiripan Judul Tugas Akhir Menggunakan Cosine Similarity	deteksi mirip judul tugas akhir guna cosine similarity

(Sumber:Albab et al., 2023)

2.1.4 *Web Scraping*

Web scraping adalah teknik yang digunakan untuk mengekstraksi data dari suatu *website* secara otomatis. Teknik ini melibatkan penargetan halaman *web* tertentu, mengambil kode *HTML* yang mendasarinya, *memparsing* data yang relevan, dan menyimpannya ke dalam sistem file atau basis data untuk dianalisis lebih lanjut (Dikilitaş et al., 2024). Dalam penelitian ini, *web scraping* digunakan untuk mengambil data judul tugas akhir dari *website* perpustakaan sebagai sumber data. Proses *scraping* dilakukan dengan memanfaatkan klasifikasi dokumen "TA TI" sebagai acuan seleksi data yang akan diambil.

2.1.5 Term Frequency–Inverse Document Frequency (TF-IDF)

Metode *Term Frequency–Inverse Document Frequency* (TF-IDF) merupakan teknik pembobotan kata yang banyak digunakan dalam pengolahan teks untuk menentukan tingkat kepentingan suatu kata dalam sebuah dokumen terhadap kumpulan dokumen lainnya. TF-IDF bekerja dengan menggabungkan dua konsep utama, yaitu *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF).

Term Frequency (TF) menggambarkan seberapa sering suatu kata muncul dalam satu dokumen. Semakin sering kata tersebut muncul, semakin tinggi nilai TF-nya. Nilai TF dapat dihitung menggunakan rumus berikut:

$$TF\ (t, d) = \ f(t, d) \qquad (2.1)$$

Keterangan:

TF = *term* (kata) yang dicari

d = dokumen

TF(t,d)= frekuensi kemunculan *term* t dalam dokumen d

F(t,d) = jumlah kemunculan *term* t dalam dokumen d.

Sementara itu, *Inverse Document Frequency* (IDF) digunakan untuk mengukur tingkat kelangkaan suatu kata pada seluruh kumpulan dokumen. Kata yang jarang muncul akan memiliki nilai IDF yang tinggi, sedangkan kata yang sering muncul pada banyak dokumen akan memiliki nilai IDF yang rendah. Rumus IDF dituliskan sebagai berikut:

$$idf(t) = \log\left(\frac{N}{df_t}\right) + 1 \qquad (2.2)$$

Keterangan:

N = Total jumlah dokumen dalam korpus (*collection*)

Dft = *Document Frequency*, yaitu jumlah dokumen yang mengandung *term* t

Nilai akhir TF-IDF diperoleh dengan mengalikan nilai TF dan IDF, sehingga menunjukkan tingkat kepentingan suatu kata pada dokumen tertentu sekaligus mempertimbangkan kelangkaannya pada dokumen lain. Perhitungan TF-IDF dirumuskan sebagai berikut:

$$TF - IDF(t,d) = TF(t,d) \times IDF(t) \qquad (2.3)$$

Menurut (Riyani et al., 2019), metode TF-IDF dikenal sebagai teknik pembobotan yang efisien, mudah diterapkan, dan akurat dalam menentukan bobot kata pada proses pengolahan teks. Sementara itu, (Meidelfi et al., 2021)menambahkan bahwa TF-IDF mampu mengubah setiap kata menjadi representasi numerik sehingga dapat digunakan dalam perhitungan kemiripan antar dokumen, termasuk dalam proses deteksi kemiripan judul tugas akhir.

2.1.6 *Cosine Similarity*

Cosine Similarity merupakan salah satu metode yang paling banyak digunakan untuk mengukur tingkat kemiripan antara dua dokumen berdasarkan representasi vektornya. Dalam pendekatan ini, setiap dokumen terlebih dahulu diubah menjadi sebuah vektor numerik, umumnya menggunakan bobot kata seperti TF-IDF. Setelah masing-masing dokumen direpresentasikan dalam bentuk vektor, nilai kemiripan dihitung berdasarkan nilai kosinus dari sudut yang terbentuk antara kedua vektor tersebut di dalam ruang multidimensi.

Metode ini bekerja dengan prinsip bahwa semakin kecil sudut antara dua vektor, semakin besar nilai kosinusnya, dan semakin mirip pula dokumen tersebut. Sebaliknya, jika sudut antara vektor semakin besar, nilai kosinus akan semakin kecil, sehingga tingkat kemiripannya juga menurun. Pendekatan ini efektif karena tidak dipengaruhi oleh panjang dokumen, tetapi lebih menekankan pada pola distribusi kata.

Secara matematis, *Cosine Similarity* antara dua vektor A dan B dapat dihitung dengan rumus:

$$Cosine\ Similarity\ (A,B) = \frac{A \cdot B}{||A|| \times ||B||} \qquad (2.4)$$

Keterangan:

A,B = Vektor representasi dokumen (hasil TF-IDF)

$A \cdot B$ = *Dot product* (perkalian titik) antara vektor A dan B

$||A||$ = Magnitude (*norma Euclidean*) vektor A

$||B||$ = Magnitude (*norma Euclidean*) vektor B

Rumus tersebut dapat dijelaskan lebih lanjut sebagai berikut:

$$A \cdot B = \sum_{i=1}^{n} A_i \times B_i \qquad (2.5)$$

$$||A|| = \sqrt{\sum_{i=1}^{n} A_i^2} \quad \text{dan} \quad ||B|| = \sqrt{\sum_{i=1}^{n} B_i^2} \qquad (2.6)$$

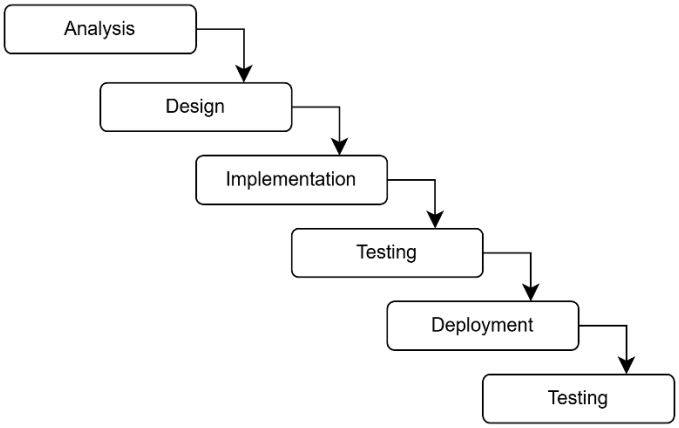
Hasil perhitungan *Cosine Similarity* berada pada rentang 0 hingga 1, di mana:

a. nilai 1 menunjukkan bahwa dua dokumen sangat mirip (arah vektornya sama),
b. nilai 0 menunjukkan tidak ada kemiripan (vektor saling tegak lurus).

Menurut (Nasrullah, 2024), algoritma *Cosine Similarity* memiliki keunggulan dalam mendeteksi tingkat kemiripan antar teks dengan tingkat akurasi yang tinggi.

2.1.7 *Waterfall*

Menurut (Fitria Nur Hasanah M.Pd, 2020) Model air terjun (*Waterfall Model*) adalah pendekatan klasik dalam pengembangan perangkat lunak yang menggambarkan metode pengembangan linier dan berurutan. Ini terdiri dari lima hingga tujuh fase, setiap fase didefinisikan oleh tugas dan tujuan yang berbeda, di mana keseluruhan fase menggambarkan siklus hidup perangkat lunak hingga pengirimannya. Setelah fase selesai, langkah pengembangan selanjutnya mengikuti dan hasil dari fase sebelumnya mengalir ke fase berikutnya.



Gambar 2.1 Metode *Waterfall*

Dikutip dari (Fitria Nur Hasanah M.Pd, 2020) ada enam tahapan dalam metode *Waterfall* sebagai berikut:

a. Analisis

Pada tahap pertama dilakukan proses pengumpulan kebutuhan secara lengkap kemudian dianalisis dan didefinisikan kebutuhan yang harus dipenuhi oleh program yang akan dibangun. Fase ini harus dikerjakan secara lengkap untuk bisa menghasilkan desain yang lengkap.

b. Desain

Pada tahap ini pengembang akan menghasilkan sebuah sistem secara keseluruhan dan menentukan alur perangkat lunak hingga algoritma yang detail.

c. Implementasi

Implementasi adalah tahapan di mana seluruh desain diubah menjadi kode program. Kode program yang dihasilkan masih berupa modul-modul yang akan diintegrasikan menjadi sistem yang lengkap.

d. Pengujian

Tahap ini dilakukan dengan menggabungkan modul-modul yang sudah dibuat dan melakukan pengujian untuk mengetahui apakah *software* yang dibuat telah sesuai dengan desainnya serta untuk memastikan apakah terdapat kesalahan pada fungsi *software* atau tidak.

e. Verifikasi

Verifikasi adalah klien atau pengguna menguji apakah sistem tersebut telah sesuai dengan yang disetujui.

f. Pemeliharaan

Pada tahap ini dilakukan proses instalasi dan perbaikan sistem sesuai yang disetujui

2.1.8 Perangkat Lunak dan Alat yang digunakan

a. Figma

Figma adalah salah satu alat desain yang digunakan untuk membuat tampilan aplikasi mobile, desktop, *website*, dan berbagai kebutuhan desain antarmuka lainnya. Figma dapat digunakan pada sistem operasi *Windows*, *Linux*, maupun macOS selama terhubung ke internet. Secara umum, Figma banyak digunakan oleh profesional di bidang *UI/UX*, *web design*, dan bidang kreatif sejenis (Muhyidin et al., 2020).

b. Draw.io

Menurut (Hendrawan et al., 2023) Draw.io adalah perangkat lunak yang digunakan untuk membuat dan mensimulasikan rancangan jaringan komputer. Perangkat lunak ini menawarkan kemampuan simulasi, penyusunan tata letak

(*layout*), serta kolaborasi sehingga memudahkan pembuatan konsep teknologi yang kompleks. Selain itu, Draw.io menyediakan berbagai komponen yang umum digunakan dalam struktur jaringan, sehingga pengguna dapat dengan mudah membuat simulasi jaringan untuk melihat cara kerja perangkat serta proses pengiriman data dari satu komputer ke komputer lainnya.

c. Javascript

JavaScript merupakan bahasa pemrograman yang digunakan dalam pengembangan *web*, di mana sebagian besar situs modern memanfaatkan JavaScript untuk menghadirkan interaktivitas pada halaman *web*. Hampir semua browser modern pada perangkat *desktop*, tablet, maupun ponsel telah menyertakan dukungan penuh terhadap *JavaScript*, menjadikannya salah satu bahasa pemrograman yang paling umum dan luas digunakan. *JavaScript* dikenal sebagai bahasa yang dinamis serta memiliki mekanisme interpretasi yang mendukung paradigma pemrograman berorientasi objek maupun pemrograman fungsional (Fandopa and Santoso, 2022).

d. Google Spreadsheets

Google Sheets merupakan aplikasi *spreadsheet* berbasis *cloud* yang dikembangkan oleh *Google* sebagai bagian dari *Google Workspace*. Aplikasi ini menyediakan struktur baris, kolom, dan cell untuk mengolah, menyimpan, serta menganalisis data secara terorganisir. Sebagai spreadsheet berbasis *web*, *Google Sheets* memungkinkan pengguna melakukan akses dan kolaborasi real-time, sehingga setiap perubahan dapat dilihat oleh seluruh pengguna yang memiliki izin akses. *Google Sheets* Memungkinkan, integrasi otomatis,

kolaborasi daring, dan kemampuan impor–ekspor menjadikan Google Sheets fleksibel untuk berbagai kebutuhan pengelolaan data (Google, 2024).

e. Python

Python adalah bahasa yang cocok untuk belajar dan sebuah pemrograman yang nyata. *Python* sangat kuat dan sebuah bahasa pemrograman tingkat tinggi yang dibuat oleh seorang programmer bernama Guido van Rossum (Halim, 2024). Proses *scraping* dilakukan menggunakan bahasa pemrograman *Python* karena memiliki berbagai *library* pendukung seperti *Requests* dan *BeautifulSoup* yang memudahkan dalam mengambil dan mem-parsing elemen-elemen pada halaman *web* perpustakaan. Dengan demikian, *Python* berperan sebagai perangkat utama dalam tahap pengumpulan data pada penelitian ini.

f. Google Drive

Google Drive adalah media penyimpanan data secara daring (*online*) berbasis internet (*cloud*) yang dikembangkan oleh *Google* dan diperkenalkan secara resmi pada tanggal 24 April 2012. Pada dasarnya layanan *Google Drive* sama seperti *cloud storage* lain seperti *DropBox* atau *OneDrive* (Suwarya, 2021). (Trilaksono et al., 2020) mengidentifikasi lima manfaat utama *Google Drive* yaitu kemampuan mengirim berbagai jenis file, sinkronisasi otomatis pada komputer, laptop dan *smartphone*, integrasi dengan aplikasi *Google* untuk kolaborasi *real-time*, pengelolaan hak akses yang fleksibel dengan pengaturan *private* atau *public*, serta penyediaan ruang penyimpanan yang besar. Dalam konteks pendidikan, (Kurniawan et al., 2023)menambahkan bahwa *Google Drive* menawarkan aksesibilitas untuk mengakses data kapan saja dan dimana

saja, skalabilitas untuk perluasan kapasitas penyimpanan tanpa perangkat tambahan, serta keamanan data yang terjamin bahkan saat terjadi kerusakan perangkat keras. *Google Drive* juga efektif sebagai solusi backup data dalam lingkungan pendidikan, memungkinkan siswa menyimpan tugas sekolah dengan aman dan berbagi file dengan guru secara terstruktur.

2.2 Penelitian Yang Relevan

Subbab ini membahas beberapa penelitian terdahulu yang berkaitan dengan topik penelitian ini. Uraian tersebut digunakan sebagai acuan dan dasar pengembangan sistem yang diusulkan, serta untuk melihat perbedaan dan keunggulan penelitian yang dilakukan saat ini.

Tabel 2.5 Penelitian Relevan

No	Penulis / Tahun	Topik Penelitian	Metode	Pembahasan	Hasil
1	(Nasrullah, 2024)	Integrasi TF-IDF dan *Cosine Similarity* untuk Deteksi Tingkat Kemiripan Judul Penelitian	TF-IDF, *Cosine Similarity, Text Preprocessing, Confusion Matrix.*	Mengembangkan aplikasi deteksi plagiarisme dokumen mahasiswa berbasis *web* dengan membandingkan dokumen pengguna terhadap database statis yang diinput secara manual.	Hasil perhitungan manual sebesar 4% dan hasil aplikasi 4,34% menunjukkan konsistensi perhitungan yang baik.
2	(Nilawati, Husaini, 2024)	Penggunaan Metode *Cosine Similarity* dan *TF-IDF* untuk Klasifikasi Judul Seminar Proposal	TF-IDF, *Cosine Similarity, Text Mining, PHP, MySQL*	Mengembangkan sistem klasifikasi judul seminar proposal untuk membantu pengelompokan judul skripsi. Proses text mining dilakukan mulai dari tahap preprocessing hingga perhitungan tingkat kemiripan dokumen, sehingga sistem mampu melakukan klasifikasi judul secara otomatis dan efisien menggunakan data latih berupa data dummy.	Sistem mampu melakukan klasifikasi judul secara otomatis dan mempercepat proses pengelompokan data.

No	Penulis / Tahun	Topik Penelitian	Metode	Pembahasan	Hasil
3	(Eko et al., 2023)	Cosine Similarity sebagai metode pendeteksi kemiripan judul dengan python	Cosine Similarity dan TF-IDF	Fokus pembahasan bertumpu pada permasalahan belum tersedianya sistem validasi judul skripsi. Sistem yang dikembangkan melakukan preprocessing teks, menghitung vektor kata, lalu menampilkan persentase kemiripan antar judul. Pendekatan ini mempermudah dosen dalam memverifikasi apakah judul tersebut sudah pernah diajukan.	Sistem yang dihasilkan mampu menampilkan persentase kemiripan judul skripsi dan berfungsi sebagai alat bantu validasi pengajuan judul. Sistem ini membantu pembimbing dalam mempercepat proses pengecekan kesamaan judul yang diajukan mahasiswa.
4	(Meidelfi et al., 2021)	Deteksi kemiripan judul tugas akhir mahasiswa TRPL PNP	*Cosine Similarity, TF-IDF, Preprocessing (Case Folding, Tokenizing, Filtering, Stemming)*	Sistem dibangun untuk membantu KBK memeriksa kesamaan judul tugas akhir yang diajukan mahasiswa dengan judul yang sudah ada. Proses meliputi preprocessing teks, pembobotan *TF-IDF*, dan perhitungan similarity menggunakan Cosine Similarity.	Sistem mampu menghitung tingkat kesamaan judul dengan cepat. Contoh perhitungan menunjukkan *similarity* 20,84% antara dua judul contoh. Sistem juga menampilkan hasil dalam persentase dan daftar judul yang mirip.
5	(Halim and Lasut, 2024)	Aplikasi Deteksi Plagiarisme Dokumen	*TF-IDF, Cosine Similarity, Text Preprocessing*	Proses dimulai dari pembersihan dan normalisasi teks, dilanjutkan dengan perhitungan bobot kata, kemudian diukur	Hasil perhitungan manual sebesar 4% dan hasil aplikasi 4,34%

No	Penulis / Tahun	Topik Penelitian	Metode	Pembahasan	Hasil
		Menggunakan Metode *TF-IDF* dan *Cosine Similarity* Berbasis *Web*	*(Lowercasing, Noise Removal, Stopword Removal, Tokenization, Feature Selection), Waterfall SDLC*	tingkat kemiripannya dengan dokumen dalam basis data. Sistem memberikan persentase kemiripan sebagai output untuk memudahkan evaluasi keaslian dokumen.	menunjukkan konsistensi perhitungan yang baik.

Analisis Tabel 2.5 menunjukkan bahwa meskipun TF-IDF dan *Cosine Similarity* telah banyak diaplikasikan, mayoritas penelitian terdahulu memiliki kelemahan dalam aspek validasi kebenaran data (*ground truth*). Evaluasi pada penelitian sebelumnya umumnya hanya membandingkan hasil perhitungan sistem dengan interpretasi subjektif peneliti secara manual, tanpa melibatkan otoritas akademik yang berwenang. Hal ini menyebabkan standar akurasi yang dihasilkan kurang merepresentasikan kebijakan institusi yang sebenarnya.

Oleh karena itu, penelitian ini mengisi celah tersebut dengan mengintegrasikan sistem secara langsung ke repositori institusi melalui *web scraping* otomatis, serta menerapkan metode evaluasi yang lebih kredibel. Penelitian ini melibatkan Koordinator Tugas Akhir sebagai validator ahli untuk menentukan ground truth terhadap kemiripan judul. Data dari pakar inilah yang kemudian digunakan untuk mengukur kinerja sistem melalui pengujian *Precision, Recall*, dan *F1-Score*. Dengan demikian, sistem yang dihasilkan tidak hanya akurat secara matematis, tetapi juga memiliki validitas yang sesuai dengan standar penilaian akademik di STMIK Palangkaraya.

BAB III
METODOLOGI PENELITIAN

3.1 Tinjauan Umum

Penelitian ini mengembangkan sistem pendeteksi kemiripan judul tugas akhir berbasis *web* yang terintegrasi langsung dengan repositori digital STMIK Palangkaraya untuk menciptakan solusi otomatis dalam pengecekan potensi kemiripan judul yang diajukan mahasiswa terhadap seluruh koleksi judul yang telah tersimpan, mengatasi keterbatasan proses manual yang saat ini masih diandalkan oleh dosen pembimbing dan koordinator tugas akhir. Sumber data penelitian diperoleh melalui mekanisme *web scraping* dari *website* perpustakaan institusi yang memuat 2316 judul tugas akhir hingga tahun 2025, menjamin kemutakhiran dataset tanpa ketergantungan pada input manual, dimana data judul kemudian diproses melalui tahapan preprocessing mencakup pembersihan dan normalisasi teks sebelum dilakukan pembobotan menggunakan metode TF-IDF dan perhitungan kemiripan dengan algoritma *Cosine Similarity* dengan pembagian tiga kategori yaitu Rendah (*similarity* di bawah 31%), Sedang (*similarity* 31-70%), dan Tinggi (similarity di atas 71%). Validasi kinerja sistem dilakukan melalui pendekatan ganda yang mengatasi keterbatasan penelitian terdahulu dengan melibatkan dosen secara langsung, dimana evaluasi akurasi algoritma mengukur metrik *Precision*, *Recall*, dan *F1-Score* terhadap *ground truth* yang ditetapkan dosen ahli, sedangkan evaluasi sistem sebagai produk dilakukan melalui uji *black-box* dengan dosen sebagai pengguna, sehingga sistem yang dibangun tidak hanya menjawab kebutuhan akurasi teknis namun juga kesiapan operasional sebagai alat bantu

keputusan yang telah divalidasi ahli dalam proses validasi judul tugas akhir di lingkungan STMIK Palangkaraya.

3.2 Jenis Penelitian

Jenis penelitian dalam tugas akhir ini menggunakan pendekatan pengembangan perangkat lunak dengan menerapkan metode Software *Development Life Cycle* (SDLC) model *Waterfall*. Model *Waterfall* dipilih karena memiliki tahapan yang terstruktur dan berurutan, sehingga sesuai dengan kebutuhan pengembangan sistem yang telah ditentukan sejak awal. Tahapan tersebut meliputi: analisis kebutuhan, perancangan sistem, implementasi, pengujian, dan pemeliharaan. Melalui pendekatan ini, proses pengembangan dilakukan secara sistematis dan terdokumentasi, sehingga menghasilkan perangkat lunak yang stabil dan sesuai tujuan penelitian.

3.3 Penelitian Yang Relevan

Metode pengembangan yang digunakan dalam pembangunan sistem pendeteksi kemiripan judul tugas akhir ini adalah *Software Development Life Cycle* (SDLC) dengan model *Waterfall*. Tahapan dalam metode ini meliputi beberapa langkah berurutan sebagai berikut:

a. Analisis Kebutuhan

   Pada tahap ini dilakukan proses pengumpulan informasi mengenai kebutuhan sistem. Analisis kebutuhan diperoleh melalui studi terhadap proses verifikasi judul yang berlangsung di STMIK Palangkaraya, studi dokumentasi terhadap *website* perpustakaan kampus, serta membaca literatur terkait teknik

pengolahan teks, TF-IDF, dan *Cosine Similarity*. Dari analisis tersebut ditentukan kebutuhan fungsional dan nonfungsional yang menjadi dasar pengembangan sistem.

b. Perancangan Sistem (*Design*)

Tahap ini mencakup pembuatan rancangan arsitektur sistem, alur proses kerja aplikasi, serta desain antarmuka pengguna. UML digunakan sebagai alat bantu pemodelan, seperti *use case diagram, sequence diagram* dan *activity diagram*. Selain itu, rancangan tampilan *website* juga disusun untuk menggambarkan struktur dan fitur utama sistem pendeteksi kemiripan judul.

c. Implementasi (*Coding*)

Tahap implementasi merupakan proses mengubah seluruh desain sistem menjadi kode program. Mengacu pada model *Waterfall*, kode program yang dibuat pada tahap ini masih berupa modul-modul terpisah yang nantinya akan diintegrasikan menjadi satu sistem yang lengkap. Pada penelitian ini, implementasi dilakukan dengan membangun modul-modul berikut:

1. Modul preprocessing teks
2. Modul pembobotan TF-IDF
3. Modul perhitungan *Cosine Similarity*
4. Modul *web scraping* untuk memperbarui dataset judul dari *website* perpustakaan
5. Modul antarmuka *website* untuk input judul dan menampilkan hasil kemiripan

d. Pengujian (*Testing*)

Pada tahap pengujian dilakukan pemeriksaan terhadap fungsi-fungsi sistem untuk memastikan bahwa setiap modul yang telah diintegrasikan berjalan sesuai dengan rancangan. Pengujian dilakukan menggunakan metode *black-box testing* dengan menguji fitur-fitur utama sistem, seperti proses pencarian judul, perhitungan kemiripan, tampilan hasil, serta proses *web scraping* dan penyimpanan data.

e. Verifikasi

Tahap verifikasi dilakukan untuk memastikan bahwa hasil penilaian kemiripan judul dari sistem sesuai dengan penilaian dosen. Verifikasi dilakukan melalui penyebaran kuesioner kepada dosen dengan membandingkan hasil *similarity* dari sistem terhadap penilaian manual dosen pada beberapa sampel judul tugas akhir. Data hasil penilaian tersebut kemudian dianalisis oleh peneliti untuk menghitung tingkat akurasi sistem menggunakan metrik *Precision*, *Recall*, dan *F1-Score*.

f. Pemeliharaan (*Maintenance*)

Tahap pemeliharaan dilakukan setelah sistem selesai dibangun dan dinyatakan berfungsi dengan baik. Pada tahap ini dilakukan perbaikan bug yang mungkin masih muncul setelah sistem digunakan. Selain itu, pemeliharaan juga mencakup proses *web scraping* secara bertahap untuk penambahan data baru pada *website* perpustakaan, sehingga sistem tetap memiliki informasi yang terbaru dan relevan.

3.4 Instrumen penelitian

Instrumen penelitian yang digunakan dalam tugas akhir ini merupakan cara yang digunakan untuk mengumpulkan data yang diperlukan dalam penelitian.

a. Studi Dokumentasi

   Studi dokumentasi dilakukan dengan mempelajari data judul tugas akhir yang tersedia pada *website* perpustakaan kampus. Informasi seperti judul, penulis, tahun, dan akses *link* digunakan sebagai dataset utama dalam proses pengembangan model pendeteksi kemiripan.

b. *Web Scraping*

   Data judul tugas akhir diperoleh melalui teknik web scraping pada website perpustakaan kampus. Data yang dikumpulkan meliputi judul, penulis, tahun, dan tautan akses, yang selanjutnya digunakan sebagai dataset utama dalam pengembangan model pendeteksi kemiripan. Teknik *web scraping* dipilih karena memungkinkan pengumpulan data secara otomatis dan efisien dari sumber yang tersedia secara daring.

3.5 Analisis Kebutuhan

Analisis kebutuhan dalam penelitian ini, yaitu perangkat keras dan perangkat lunak yang dibutuhkan untuk mendukung kelancaran pelaksanaan dan pengolahan data.

a. Perangkat Keras (*Hardware*)

   Perangkat keras merupakan komponen fisik yang digunakan untuk mendukung proses pengembangan dan pengujian sistem dalam penelitian ini. Spesifikasi perangkat keras yang digunakan disesuaikan dengan kebutuhan

pemrosesan data dan pengoperasian sistem pendeteksi kemiripan judul. Rincian perangkat keras yang digunakan ditunjukkan pada Tabel 3.1.

Tabel 3.1 Perangkat Keras

No	Perangkat Keras	Spesifikasi
1.	Laptop	Lenovo Ideapad 5
2.	Processor	AMD Ryzen 4500
3.	RAM	8.00 GB
4.	Storage	512 Gb

b. Perangkat Lunak (*Software*)

Perangkat lunak merupakan komponen pendukung yang digunakan dalam proses pengembangan, pengujian, dan pengoperasian sistem pendeteksi kemiripan judul tugas akhir. Perangkat lunak yang digunakan meliputi sistem operasi, bahasa pemrograman, serta *tools* pendukung lainnya. Rincian perangkat lunak yang digunakan ditunjukkan pada Tabel 3.2.

Tabel 3.2 Perangkat Lunak

No	Perangkat Lunak	Keterangan
1.	Windows 11	Sebagai Sistem Operasi
2.	Javascript	Sebagai bahasa pemrograman
3.	Google Spreadsheets	Sebagai *Database*
4.	Google Drive	Penyimpanan *Cloud*
4.	Python 3	Sebagai *web scraping*
5.	Figma	Sebagai perancangan (*interface*)

3.5.1 Analisis Data

Dalam melakukan analisis data diperlukan data yang valid dan dapat dipertanggungjawabkan. Oleh karena itu, penulis melakukan analisis berdasarkan hasil perhitungan menggunakan rumus yang digunakan serta kajian referensi resmi. Analisis data pada penelitian ini dilakukan melalui perhitungan manual untuk

memverifikasi hasil sistem. Kategori kemiripan judul ditunjukkan pada Tabel 3.3. merujuk pada penelitian yang dilakukan oleh (Lumbansiantar et al., 2023).

Tabel 3.3 Tingkat Kemiripan Judul

Rentang Nilai	Kategori	Interpretasi
0-31%	Rendah	Judul memiliki perbedaan signifikan dalam metode, objek, atau fokus penelitian. Judul dapat digunakan.
31-70%	Sedang	Judul memiliki kesamaan dalam beberapa aspek seperti metode atau objek penelitian. Perlu pengecekan lebih lanjut dan klarifikasi perbedaan.
71-100%	Tinggi	Judul memiliki kesamaan yang sangat signifikan. Perlu revisi judul untuk memastikan originalitas penelitian.

Contoh soal

Q : Penerapan Metode Fuzzy Mamdani Untuk Diagnosa Penyakit Pata Tanaman Peria

D1 : Sistem Pakar Diagnosa Penyakit Pada Tanaman Jeruk Siam Banjar Menggunakan Metode Fuzzy Mamdani

D2 : Sistem Pakar Diagnosa Hama Dan Penyakit Pada Tanaman Sukun Menggunakan Metode Certainty Factor

a. *Preprocessing*

1. *Lowercasing*

Tahap pertama pada proses *preprocessing* adalah mengubah seluruh judul tugas akhir menjadi huruf kecil (*lowercase*) untuk menyeragamkan bentuk data. Proses ini bertujuan agar sistem tidak membedakan antara huruf

besar dan huruf kecil dalam tahap analisis selanjutnya. Hasil proses *case folding* pada judul tugas akhir ditunjukkan pada Tabel 3.4.

Tabel 3.4 Hasil *Case Folding* pada Judul Tugas Akhir

Teks Asli	Hasil
Penerapan Metode Fuzzy Mamdani Untuk Diagnosa Penyakit Pata Tanaman Peria	penerapan metode fuzzy mamdani untuk diagnosa penyakit pata tanaman peria
Sistem Pakar Diagnosa Penyakit Pada Tanaman Jeruk Siam Banjar Menggunakan Metode Fuzzy Mamdani	sistem pakar diagnosa penyakit pada tanaman jeruk siam banjar menggunakan metode fuzzy mamdani
Sistem Pakar Diagnosa Hama Dan Penyakit Pada Tanaman Sukun Menggunakan Metode Certainty Factor	sistem pakar diagnosa hama dan penyakit pada tanaman sukun menggunakan metode certainty factor

2. *Tokenizing*

Setelah dilakukan proses *lowercasing*, setiap judul tugas akhir kemudian dipecah menjadi token atau kata per kata. Proses ini bertujuan agar setiap kata dapat dianalisis secara individual pada tahap selanjutnya. Hasil proses *tokenizing* pada judul tugas akhir ditunjukkan pada Tabel 3.5.

Tabel 3.5 Hasil *Tokenizing* pada Judul Tugas Akhir

Token	Hasil
penerapan metode fuzzy mamdani untuk diagnosa penyakit pata tanaman peria	[penerapan, metode, fuzzy, mamdani, untuk, diagnosa, penyakit, pata, tanaman, peria]
sistem pakar diagnosa penyakit pada tanaman jeruk siam banjar menggunakan metode fuzzy mamdani	[sistem, pakar, diagnosa, penyakit, pada, tanaman, jeruk, siam, banjar, menggunakan, metode, fuzzy, mamdani]
sistem pakar diagnosa hama dan penyakit pada tanaman sukun	[sistem, pakar, diagnosa, hama, dan, penyakit, pada, tanaman,

Token	Hasil
menggunakan metode certainty factor	sukun, menggunakan, metode, certainty, factor]

3. *Stopwords Removal*

Pada tahap ini, token yang mengandung kata-kata umum (*stopwords*) dihapus agar hanya kata-kata yang relevan dengan konteks judul yang dipertahankan. Proses ini bertujuan untuk meningkatkan kualitas analisis kemiripan. Hasil proses *stopword removal* pada judul tugas akhir ditunjukkan pada Tabel 3.6.

Tabel 3.6 Hasil *Stopword Removal* pada Judul Tugas Akhir

Tokens	Hasil
[penerapan, metode, fuzzy, mamdani, untuk, diagnosa, penyakit, pata, tanaman, peria]	[penerapan, metode, fuzzy, mamdani, diagnosa, penyakit, pata, tanaman, peria]
[sistem, pakar, diagnosa, penyakit, pada, tanaman, jeruk, siam, banjar, menggunakan, metode, fuzzy, mamdani]	[sistem, pakar, diagnosa, penyakit, tanaman, jeruk, siam, banjar, metode, fuzzy, mamdani]
[sistem, pakar, diagnosa, hama, dan, penyakit, pada, tanaman, sukun, menggunakan, metode, certainty, factor]	[sistem, pakar, diagnosa, hama, penyakit, tanaman, sukun, metode, certainty, factor]

4. *Stemming*

Setelah proses *stopword removal* dilakukan, token yang tersisa kemudian diproses untuk me mperoleh bentuk kata dasarnya melalui proses stemming. Tahap ini bertujuan menyederhanakan kata berimbuhan menjadi kata dasar agar perhitungan TF-IDF menjadi lebih

akurat. Hasil proses stemming pada judul tugas akhir ditunjukkan pada Tabel 3.7.

Tabel 3.7 Hasil *Stemming* pada Judul Tugas Akhir

Tokens	Hasil
[penerapan, metode, fuzzy, mamdani, diagnosa, penyakit, pata, tanaman, peria]	[terap, metode, fuzzy, mamdani, diagnosa, sakit, pata, tanam, peria]
[sistem, pakar, diagnosa, penyakit, tanaman, jeruk, siam, banjar, metode, fuzzy, mamdani]	[sistem, pakar, diagnosa, sakit, tanam, jeruk, siam, banjar, metode, fuzzy, mamdani]
[sistem, pakar, diagnosa, hama, penyakit, tanaman, sukun, metode, certainty, factor]	[sistem, pakar, diagnosa, hama, sakit, tanam, sukun, metode, certainty, factor]

b. TF

Tahap Term Frequency (TF) digunakan untuk menghitung frekuensi kemunculan setiap term pada masing-masing dokumen judul tugas akhir. Nilai TF menunjukkan seberapa sering suatu kata muncul dalam satu dokumen dan menjadi dasar dalam perhitungan bobot TF-IDF. Frekuensi kemunculan term pada setiap dokumen dengan rumus persamaan (2.1) ditunjukkan pada Tabel 3.8.

Tabel 3.8 Frekuensi Kemunculan *Term* (TF) pada Setiap Dokumen

Tokens	TF (Q)	TF (D1)	TF (D2)
terap	1	0	0
metode	1	1	1
fuzzy	1	1	0
mamdani	1	1	0
diagnosa	1	1	1
sakit	1	1	1
pata	1	0	0
tanam	1	1	1
peria	1	0	0
sistem	0	1	1

Tokens	TF (Q)	TF (D1)	TF (D2)
pakar	0	1	1
jeruk	0	1	0
siam	0	1	0
banjar	0	1	0
hama	0	0	1
sukun	0	0	1
certainty	0	0	1
factor	0	0	1

c. IDF

Tahap *Inverse Document Frequency* (IDF) digunakan untuk mengukur tingkat kepentingan suatu term dalam keseluruhan dokumen. Nilai IDF akan semakin besar jika suatu term jarang muncul pada dokumen lain, sehingga memberikan bobot lebih pada kata yang bersifat spesifik. Hasil perhitungan nilai IDF untuk setiap term dihitung dengan persamaan (2.2) ditunjukkan pada Tabel 3.9.

Tabel 3.9 Perhitungan Nilai IDF untuk Setiap Term

Tokens	*DF*	*IDF*
terap	1	1.477
metode	3	1
fuzzy	2	1.176
mamdani	2	1.176
diagnosa	3	1
sakit	3	1
pata	1	1.477
tanam	3	1
peria	1	1.477
sistem	2	1.176
pakar	2	1.176
jeruk	1	1.477
siam	1	1.477
banjar	1	1.477
hama	1	1.477
sukun	1	1.477

Tokens	*DF*	*IDF*
certainty	1	1.477
factor	1	1.477

d. TF-IDF

Tahap TF-IDF merupakan proses pembobotan term dengan mengombinasikan nilai *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF). Pembobotan ini bertujuan untuk menonjolkan kata-kata yang penting dalam suatu dokumen serta mengurangi pengaruh kata yang sering muncul pada banyak dokumen. Nilai bobot TF-IDF setiap term pada masing-masing dokumen ditunjukkan pada Tabel 3.10.

Tabel 3.10 Nilai Bobot TF-IDF Setiap *Term* pada Masing-Masing Dokumen

Tokens	Q	D1	D2
terap	1.477	0	0
metode	1	1	1
fuzzy	1.176	1.176	0
mamdani	1.176	1.176	0
diagnosa	1	1	1
sakit	1	1	1
pata	1.477	0	0
tanam	1	1	1
peria	1.477	0	0
sistem	0	1.176	1.176
pakar	0	1.176	1.176
jeruk	0	1.477	0
siam	0	1.477	0
banjar	0	1.477	0
hama	0	0	1.477
sukun	0	0	1.477
certainty	0	0	1.477
factor	0	0	1.477

e. *Cosine Similarity*

Proses perhitungan *Cosine Similarity* dapat dihitung dengan menghitung dot product terlebih dahulu dengan persamaan (2.5).

Dot Product A·B

metode: 1 × 1 = 1

fuzzy: 1.176 × 1.176 = 1.383

mamdani: 1.176 × 1.176 = 1.383

diagnosa: 1 × 1 = 1

sakit: 1 × 1 = 1

tanam: 1 × 1 = 1

Total *Dot Product* = 6.766

Persamaan (2.6) digunakan untuk menghitung panjang vektor.

Magnitude ||Q|| =

$$\sqrt{1{,}477^2 + 1^2 + 1{,}176^2 + 1{,}176^2 + 1^2 + 1^2 + 1{,}477^2 + 1^2 + 1{,}477^2} = 3.648$$

Magnitude ||D1|| =

$$\sqrt{1^2 + 1{,}176^2 + 1{,}176^2 + 1^2 + 1^2 + 1^2 + 1{,}176^2 + 1{,}176^2 + 1{,}477^2 + 1{,}477^2 + 1{,}477^2}$$

=4,009

Cosine Similarity (Q, D1) = $\frac{6{,}766}{(3{,}648 \times 4{,}009)} = 0{,}4626 = 46{,}3\%$

Dot Product A·B antara Q dan D2

metode: 1 × 1 = 1

diagnosa: 1 × 1 = 1

sakit: 1 × 1 = 1

tanam: 1 × 1 = 1

Total Dot Product = 4

Magnitude ||D2|| =

$$\sqrt{1^2 + 1^2 + 1^2 + 1^2 + 1{,}176^2 + 1{,}176^2 + 1{,}477^2 + 1{,}477^2 + 1{,}477^2 + 1{,}477^2}$$

=3,936

Persamaan (2.4) kembali digunakan untuk menghitung *Cosine Similarity*.

Tahap *Cosine Similarity* digunakan untuk mengukur tingkat kemiripan antar judul tugas akhir berdasarkan vektor bobot TF-IDF. Hasil nilai *Cosine Similarity* berada pada rentang 0 hingga 1, di mana nilai yang semakin mendekati 1 menunjukkan tingkat kemiripan yang semakin tinggi. Hasil akhir perhitungan *Cosine Similarity* ditunjukkan pada Tabel 3.11.

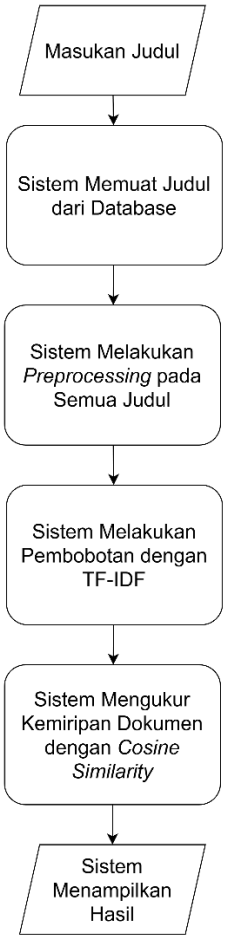
Tabel 3.11 Hasil Akhir Cosine Similarity

Dokumen	*Cosine Similarity*	Persentase Kemiripan	Kategori
D1	0.463	46.3%	Sedang
D2	0.279	27.9%	Rendah

3.5.2 Analisis Proses

Analisis proses pada sistem ini dibagi menjadi dua bagian, yaitu alur kerja website dan alur kerja *web scraping*. Alur kerja website menjelaskan bagaimana sistem memproses input judul hingga menghasilkan nilai kemiripan, sedangkan alur kerja *web scraping* menggambarkan proses pengambilan dan pembaruan data judul dari website perpustakaan. Kedua proses ini saling mendukung dalam menghasilkan sistem yang berfungsi secara optimal.

a. Proses Alur Kerja *Website*

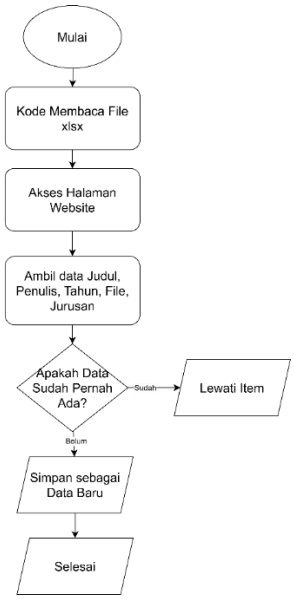


Gambar 3.1 Flowchart Alur Kerja *Website*

Proses alur kerja sistem pendeteksi kemiripan judul dimulai ketika pengguna memasukkan judul Tugas Akhir ke dalam form pada *website*. Setelah menerima input tersebut, sistem memuat seluruh judul yang tersimpan pada *database* digital untuk dijadikan pembanding. Setiap judul, termasuk judul input pengguna, kemudian diproses melalui tahap *preprocessing* yang

meliputi *case folding*, *tokenizing*, *stopword removal*, dan *stemming* agar seluruh teks berada dalam bentuk yang seragam. Selanjutnya, sistem menghitung bobot setiap kata menggunakan metode TF-IDF untuk mengubah setiap judul menjadi vektor numerik. Berdasarkan vektor tersebut, tingkat kemiripan antara judul pengguna dan seluruh judul pada database dihitung menggunakan algoritma *Cosine Similarity*. Pada tahap akhir, sistem menampilkan hasil berupa persentase kemiripan dan daftar judul yang paling relevan, sehingga pengguna dapat mengetahui apakah judul yang diajukan memiliki tingkat kesamaan dengan penelitian sebelumnya.

b. Proses Alur Kerja *Web Scraping*



Gambar 3.2 *Flowchart* Alur Kerja *Scraping*

Proses *web scraping* dilakukan untuk mengambil data judul Tugas Akhir secara otomatis dari website Perpustakaan STMIK Palangkaraya. Sistem memuat data lama untuk mencegah duplikasi, lalu mengekstraksi informasi seperti judul,

penulis, tahun, NIM, jurusan, dan tautan file menggunakan library BeautifulSoup. Data baru yang ditemukan kemudian disimpan ke dalam file Excel sebagai *database* utama *website* pendeteksi kemiripan judul.

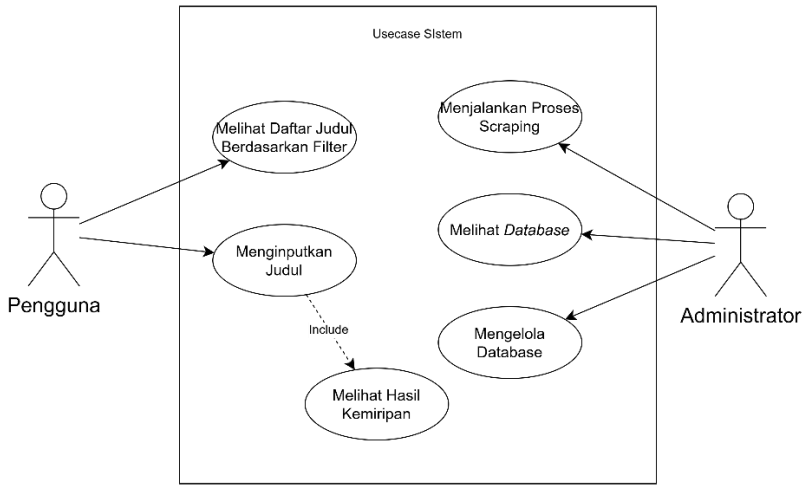
3.5.3 Analisis Kelemahan

Salah satu kelemahan utama dalam sistem ini adalah proses pengambilan data melalui *web scraping* masih perlu dijalankan setiap kali data ingin diperbarui, meskipun penyimpanan hasil *scraping* berjalan otomatis.

3.6 Desain

3.6.1 Desain Proses

a. *Usecase Diagram*

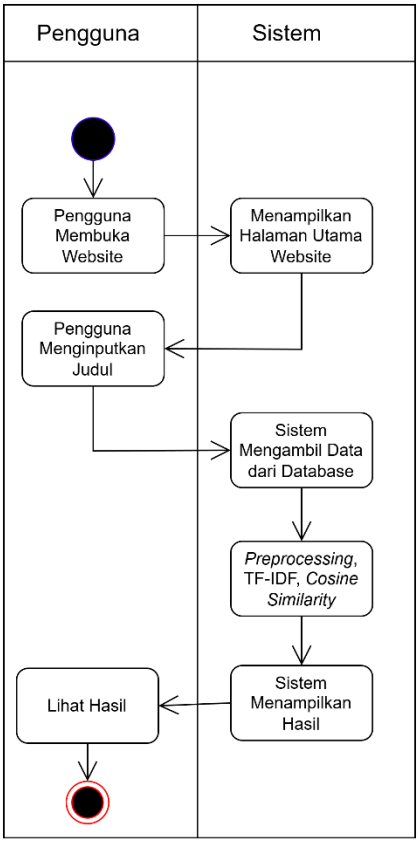


Gambar 3.3 *Usecase Diagram* proses

Pada sistem pendeteksi kemiripan judul ini terdapat dua aktor, yaitu Pengguna dan Administrator. Pengguna dapat melihat daftar judul berdasarkan filter, serta menginputkan judul yang ingin diperiksa. Setelah judul dimasukkan, sistem secara otomatis menampilkan hasil kemiripan

sebagai bagian dari proses tersebut. Sementara itu, Administrator bertanggung jawab dalam menjaga ketersediaan dan pembaruan data melalui proses *web scraping*, pengelolaan isi basis data, serta pemantauan data yang telah tersimpan.

b. *Activity Diagram*



Gambar 3.4 *Activity Diagram* Alur Sistem

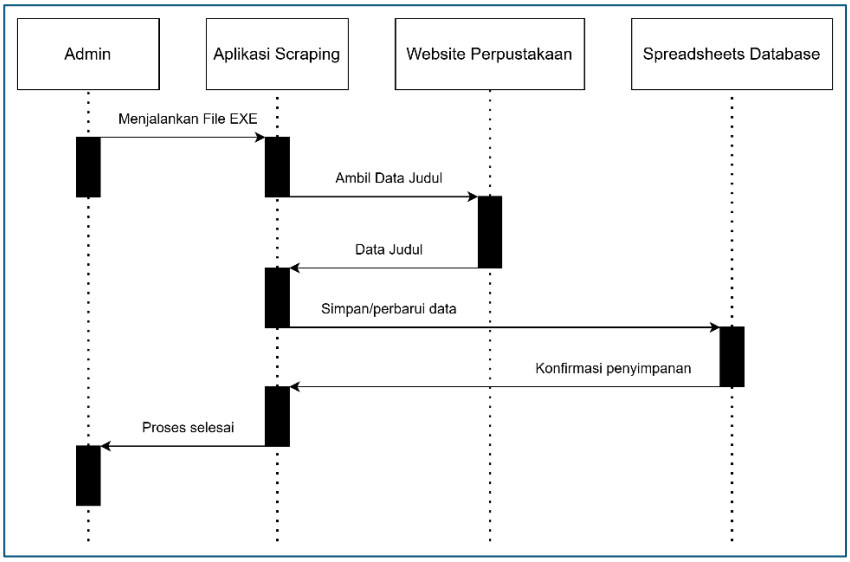
Activity diagram menggambarkan alur proses ketika pengguna melakukan pengecekan kemiripan judul melalui *website*. Proses dimulai saat pengguna membuka halaman utama *website*, kemudian sistem menampilkan antarmuka awal yang berisi *form input* judul. Setelah halaman tersebut

ditampilkan, pengguna memasukkan judul Tugas Akhir yang ingin diperiksa.

Setelah judul diterima, sistem mengambil seluruh data judul pembanding dari database untuk digunakan dalam proses perhitungan. Selanjutnya, sistem melakukan serangkaian proses pemrosesan teks, seperti *Case Folding*, *Tokenizing*, *Stopwords Removal* dan *Stemming*, perhitungan TF-IDF, dan *Cosine Similarity* untuk menentukan tingkat kemiripan judul yang diinputkan pengguna dengan judul-judul yang tersedia dalam database. Setelah perhitungan selesai, sistem menyajikan hasil berupa persentase kemiripan dan daftar judul terkait kepada pengguna. Proses berakhir ketika pengguna melihat hasil yang ditampilkan oleh sistem.

c. Sequence Diagram

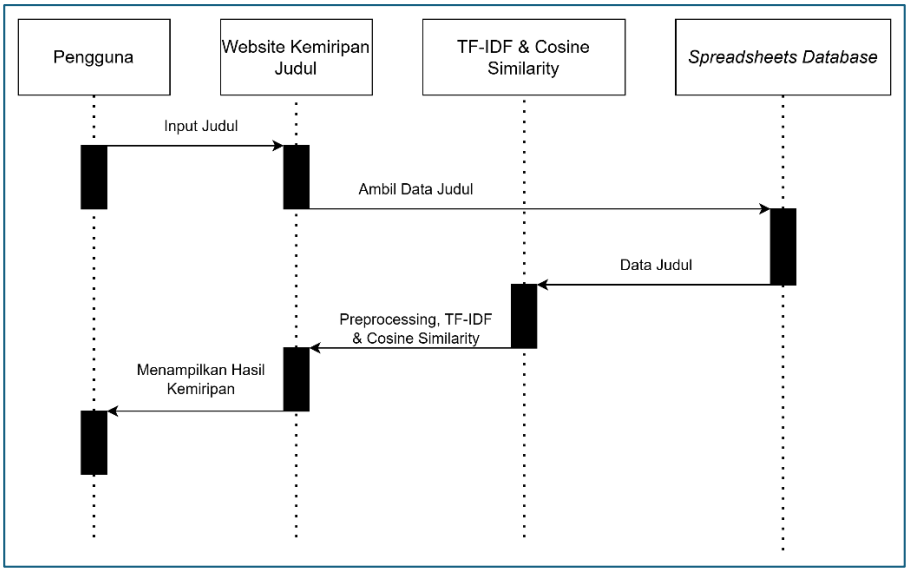
1) Admin



Gambar 3.5 *Sequence Diagram* Admin

Sequence Diagram Admin menggambarkan alur interaksi antara admin dengan sistem dalam proses pembaruan data judul tugas akhir. Proses diawali ketika admin menjalankan aplikasi *web scraping* melalui file berekstensi .exe. Selanjutnya, aplikasi scraping akan secara otomatis mengakses *website* perpustakaan STMIK Palangkaraya untuk mengambil data judul tugas akhir terbaru. Setelah proses pengambilan data selesai, aplikasi akan melakukan penyimpanan atau pembaruan data ke dalam *database spreadsheet* yang digunakan oleh sistem deteksi kemiripan judul. Setelah data berhasil diperbarui, sistem memberikan informasi bahwa proses *scraping* telah selesai dilaksanakan. Dengan adanya proses ini, data judul tugas akhir yang digunakan oleh sistem akan selalu terbarui tanpa perlu dilakukan input data secara manual oleh admin.

2) Pengguna

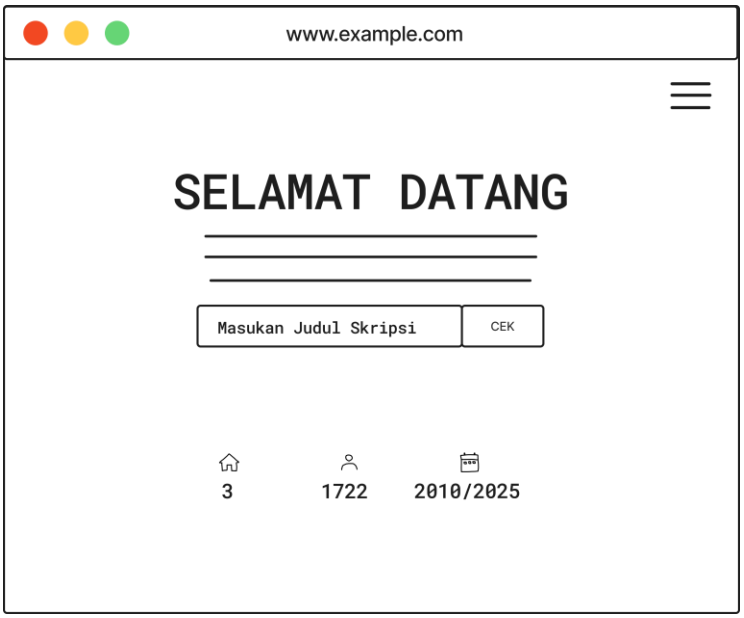


Gambar 3.6 *Sequence Diagram* Pengguna

Sequence Diagram Pengguna menggambarkan alur interaksi antara pengguna dengan sistem dalam proses deteksi kemiripan judul tugas akhir. Proses dimulai ketika pengguna memasukkan judul tugas akhir yang akan diperiksa melalui halaman *website*. Selanjutnya, sistem mengambil data judul tugas akhir yang tersimpan pada *database spreadsheet*. Setelah data berhasil diperoleh, sistem melakukan proses deteksi kemiripan yang meliputi tahapan preprocessing, perhitungan bobot menggunakan metode TF-IDF, dan perhitungan tingkat kemiripan menggunakan metode *Cosine Similarity*. Hasil perhitungan kemudian diklasifikasikan berdasarkan tingkat kemiripannya dan ditampilkan kepada pengguna dalam bentuk persentase kemiripan beserta kategori yang sesuai.

3.6.2 Desain Perangkat

a. Halaman Utama

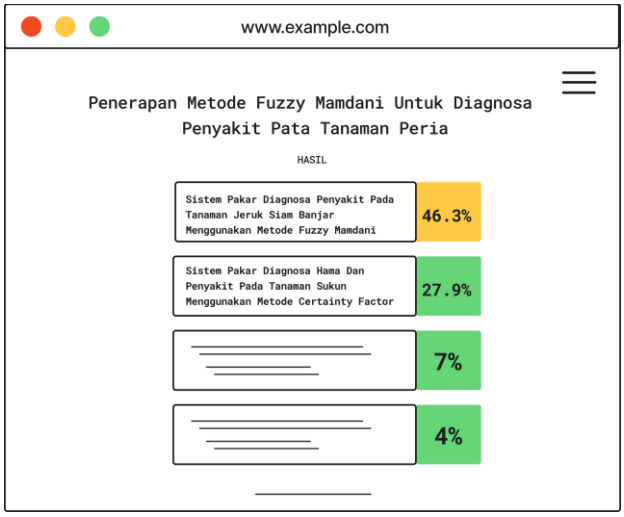


Gambar 3.7 Desain Halaman Utama

Halaman utama dirancang sebagai halaman pertama yang dilihat oleh pengguna saat mengakses sistem. Pada bagian atas halaman ditampilkan tulisan "Selamat Datang" sebagai sapaan awal, diikuti oleh sub judul yang menjelaskan tujuan sistem, yaitu membantu pengguna mengecek kemiripan judul tugas akhir. Di bagian tengah halaman tersedia kolom input judul tempat pengguna memasukkan judul yang ingin diperiksa, disertai dengan tombol "Cari" untuk memulai proses pengecekan.

Selain itu, halaman utama juga menampilkan sejumlah statistik ringkas yang memberikan gambaran mengenai isi database sistem, seperti jumlah total program studi yang terdata, jumlah keseluruhan judul tugas akhir yang tersedia, serta rentang tahun dari data yang terkumpul. Informasi ini ditampilkan dalam bentuk kartu ringkas untuk memberikan konteks kepada pengguna mengenai cakupan data yang digunakan dalam proses pendeteksian kemiripan.

b. Halaman Hasil

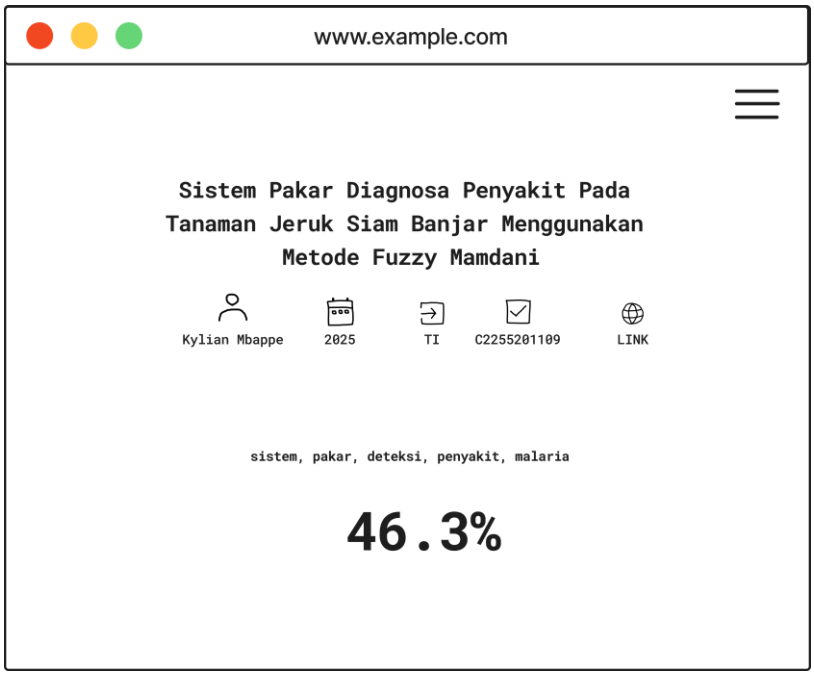


Gambar 3.8 Desain Halaman Hasil

Halaman hasil merupakan bagian yang menampilkan keluaran dari proses pengecekan kemiripan judul. Setelah pengguna memasukkan judul dan menekan tombol “Cari”, sistem akan menampilkan daftar judul yang paling mirip dengan judul yang dimasukkan. Daftar ini disajikan dalam bentuk list berisi maksimal 10 judul yang memiliki tingkat kemiripan tertinggi berdasarkan perhitungan TF-IDF dan *Cosine Similarity*.

Setiap judul dalam daftar ditampilkan bersama persentase kemiripan, sehingga pengguna dapat langsung mengidentifikasi apakah judul mereka tergolong aman atau memiliki kemiripan yang tinggi dan perlu direvisi.

c. Halaman Detail

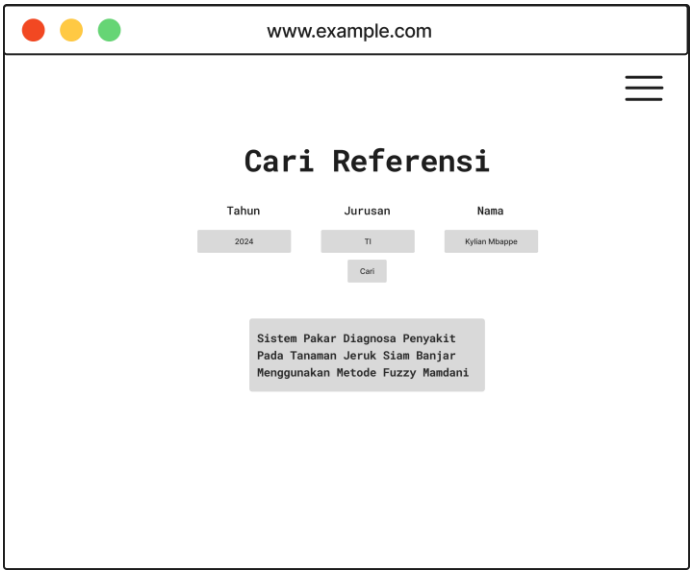


Gambar 3.9 Desain Halaman Detail

Halaman detail dirancang untuk menampilkan informasi lengkap dari salah satu judul yang dianggap mirip berdasarkan hasil perhitungan sistem. Pada halaman ini, pengguna dapat melihat judul tugas akhir yang serupa beserta informasi penting lainnya seperti nama penulis, tahun publikasi, program studi, dan NIM. Selain itu, halaman ini juga menyediakan tautan langsung ke halaman perpustakaan kampus yang diperoleh dari proses *web scraping*, sehingga pengguna dapat melihat sumber aslinya secara lebih lengkap.

Sistem juga menampilkan persentase kemiripan antara judul yang diperiksa dengan judul yang ditampilkan, serta daftar kata-kata yang sama atau memiliki kesesuaian tinggi yang berkontribusi pada nilai *similarity* tersebut. Dengan tampilan detail ini, pengguna dapat memahami secara jelas faktor apa yang membuat kedua judul serupa dan menentukan apakah judul yang mereka ajukan perlu diperbaiki atau diganti.

d. Halaman Referensi



Gambar 3.10 Desain Halaman Daftar Judul

Halaman daftar judul berfungsi sebagai pusat pencarian referensi yang menampilkan seluruh judul tugas akhir pada *database*. Pengguna dapat melakukan pencarian dan penyaringan berdasarkan tahun, jurusan, atau nama penulis. Hasil filter ditampilkan dalam bentuk daftar, dan setiap judul dapat dibuka untuk melihat informasi detail. Halaman ini memudahkan pengguna menelusuri referensi secara cepat dan efisien.

3.6.3 Desain Basisdata

Basis data pada sistem ini menggunakan *Google Spreadsheets* sebagai media penyimpanan utama yang menampung seluruh data judul tugas akhir hasil proses *web scraping* dari *website* Perpustakaan STMIK Palangkaraya. Struktur tabel dirancang secara sederhana namun telah mencakup seluruh informasi yang dibutuhkan dalam proses pendeteksian kemiripan judul. Setiap baris pada tabel

merepresentasikan satu entri judul tugas akhir, sedangkan setiap kolom berisi atribut yang relevan. Desain kolom pada basis data ditunjukkan pada Tabel 3.12.

Tabel 3.12 Database Spreadsheets

Nama Kolom	Tipe Data	Keterangan
ID	INT	Berisikan ID judul
Judul	TEXT	Judul Tugas Akhir
NIM	VARCHAR	Nama Penulis/Pengarang
Tahun	INT	NIM Mahasiswa
Penulis	VARCHAR	Tahun Penyusunan
Jurusan	VARCHAR	TI/SI/MI
File	LINK	Link Halaman Perpustakaan Hasil Scraping

BAB IV
PEMBAHASAN

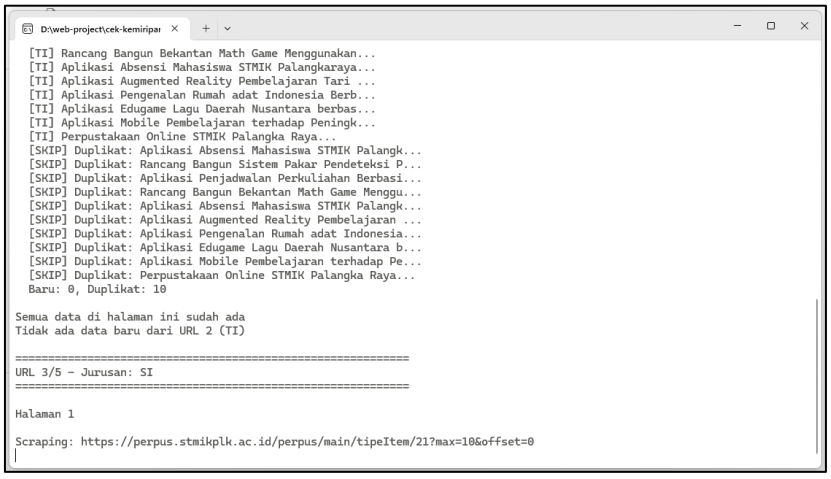
4.1 Hasil

Hasil penelitian ini menghasilkan dua sistem yang saling mendukung, yaitu aplikasi berbasis *website* sebagai media penyajian data, serta program *Python* yang berfungsi untuk membantu dalam penambahan data pada sistem kemiripan.

4.1.1 Implementasi Sistem

a. Sistem *Scraping*

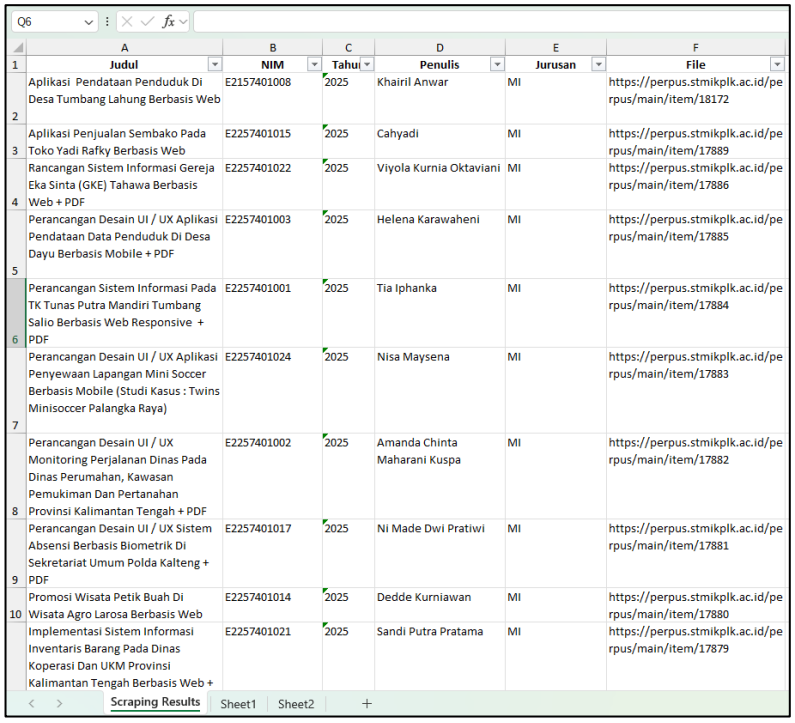
Sistem *scraping* dikemas dalam bentuk file *executable* (.exe) sehingga dapat dijalankan langsung tanpa memerlukan instalasi *Python* terlebih dahulu. Proses *scraping* dimulai ketika pengguna mengklik file *executable* tersebut, kemudian sistem secara otomatis akan mengakses *website* Perpustakaan STMIK Palangkaraya dan mulai mengambil data judul tugas akhir yang tersedia.



Gambar 4.1 Proses Scraping

Proses scraping berhasil mengumpulkan sebanyak 2340 data judul tugas akhir yang mencakup atribut judul, nama penulis, NIM, jurusan, tahun, dan tautan

file dokumen. Sistem scraping juga dilengkapi dengan mekanisme pengecekan duplikasi, sehingga data yang sudah ada sebelumnya tidak akan tersimpan ganda ke dalam *database*.

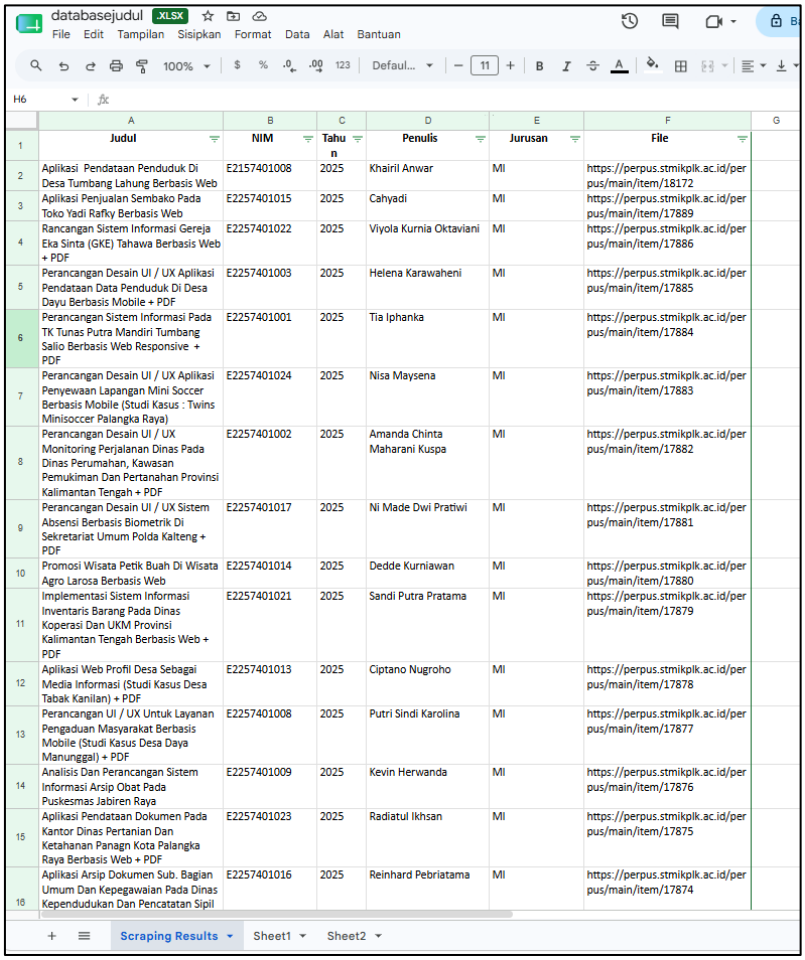


	A	B	C	D	E	F
1	Judul	NIM	Tahu	Penulis	Jurusan	File
2	Aplikasi Pendataan Penduduk Di Desa Tumbang Lahung Berbasis Web	E2157401008	2025	Khairil Anwar	MI	https://perpus.stmikplk.ac.id/perpus/main/item/18172
3	Aplikasi Penjualan Sembako Pada Toko Yadi Rafky Berbasis Web	E2257401015	2025	Cahyadi	MI	https://perpus.stmikplk.ac.id/perpus/main/item/17889
4	Rancangan Sistem Informasi Gereja Eka Sinta (GKE) Tahawa Berbasis Web + PDF	E2257401022	2025	Viyola Kurnia Oktaviani	MI	https://perpus.stmikplk.ac.id/perpus/main/item/17886
5	Perancangan Desain UI / UX Aplikasi Pendataan Data Penduduk Di Desa Dayu Berbasis Mobile + PDF	E2257401003	2025	Helena Karawaheni	MI	https://perpus.stmikplk.ac.id/perpus/main/item/17885
6	Perancangan Sistem Informasi Pada TK Tunas Putra Mandiri Tumbang Salio Berbasis Web Responsive + PDF	E2257401001	2025	Tia Iphanka	MI	https://perpus.stmikplk.ac.id/perpus/main/item/17884
7	Perancangan Desain UI / UX Aplikasi Penyewaan Lapangan Mini Soccer Berbasis Mobile (Studi Kasus : Twins Minisoccer Palangka Raya)	E2257401024	2025	Nisa Maysena	MI	https://perpus.stmikplk.ac.id/perpus/main/item/17883
8	Perancangan Desain UI / UX Monitoring Perjalanan Dinas Pada Dinas Perumahan, Kawasan Pemukiman Dan Pertanahan Provinsi Kalimantan Tengah + PDF	E2257401002	2025	Amanda Chinta Maharani Kuspa	MI	https://perpus.stmikplk.ac.id/perpus/main/item/17882
9	Perancangan Desain UI / UX Sistem Absensi Berbasis Biometrik Di Sekretariat Umum Polda Kalteng + PDF	E2257401017	2025	Ni Made Dwi Pratiwi	MI	https://perpus.stmikplk.ac.id/perpus/main/item/17881
10	Promosi Wisata Petik Buah Di Wisata Agro Larosa Berbasis Web	E2257401014	2025	Dedde Kurniawan	MI	https://perpus.stmikplk.ac.id/perpus/main/item/17880
	Implementasi Sistem Informasi Inventaris Barang Pada Dinas Koperasi Dan UKM Provinsi Kalimantan Tengah Berbasis Web +	E2257401021	2025	Sandi Putra Pratama	MI	https://perpus.stmikplk.ac.id/perpus/main/item/17879

Scraping Results | Sheet1 | Sheet2

Gambar 4.2 Hasil *Scraping* Data Judul

File *Excel* hasil *scraping* kemudian diunggah ke *Google Drive* dan dibuka menggunakan *Google Spreadsheets*. Penggunaan *Google Spreadsheets* dipilih karena memungkinkan data diakses secara daring tanpa memerlukan *server database* khusus, sehingga lebih praktis dan efisien untuk kebutuhan sistem ini. Pada *Google Spreadsheets*, data disimpan dalam beberapa halaman *sheet* yang terpisah, yaitu sheet untuk data judul tugas akhir, *sheet* untuk daftar *stopword*, dan *sheet* untuk kamus koreksi *stemming*.



databasejudul .XLSX

File Edit Tampilan Sisipkan Format Data Alat Bantuan

	Judul	NIM	Tahun	Penulis	Jurusan	File
2	Aplikasi Pendataan Penduduk Di Desa Tumbang Lahung Berbasis Web	E2157401008	2025	Khairil Anwar	MI	https://perpus.stmikplk.ac.id/perpus/main/item/18172
3	Aplikasi Penjualan Sembako Pada Toko Yadi Rafky Berbasis Web	E2257401015	2025	Cahyadi	MI	https://perpus.stmikplk.ac.id/perpus/main/item/17889
4	Rancangan Sistem Informasi Gereja Eka Sinta (GKE) Tahawa Berbasis Web + PDF	E2257401022	2025	Viyola Kurnia Oktaviani	MI	https://perpus.stmikplk.ac.id/perpus/main/item/17886
5	Perancangan Desain UI / UX Aplikasi Pendataan Data Penduduk Di Desa Dayu Berbasis Mobile + PDF	E2257401003	2025	Helena Karawaheni	MI	https://perpus.stmikplk.ac.id/perpus/main/item/17885
6	Perancangan Sistem Informasi Pada TK Tunas Putra Mandiri Tumbang Sallo Berbasis Web Responsive + PDF	E2257401001	2025	Tia Iphanka	MI	https://perpus.stmikplk.ac.id/perpus/main/item/17884
7	Perancangan Desain UI / UX Aplikasi Penyewaan Lapangan Mini Soccer Berbasis Mobile (Studi Kasus : Twins Minisoccer Palangka Raya)	E2257401024	2025	Nisa Maysena	MI	https://perpus.stmikplk.ac.id/perpus/main/item/17883
8	Perancangan Desain UI / UX Monitoring Perjalanan Dinas Pada Dinas Perumahan, Kawasan Pemukiman Dan Pertanahan Provinsi Kalimantan Tengah + PDF	E2257401002	2025	Amanda Chinta Maharani Kuspa	MI	https://perpus.stmikplk.ac.id/perpus/main/item/17882
9	Perancangan Desain UI / UX Sistem Absensi Berbasis Biometrik Di Sekretariat Umum Polda Kalteng + PDF	E2257401017	2025	Ni Made Dwi Pratiwi	MI	https://perpus.stmikplk.ac.id/perpus/main/item/17881
10	Promosi Wisata Petik Buah Di Wisata Agro Larosa Berbasis Web	E2257401014	2025	Dedde Kurniawan	MI	https://perpus.stmikplk.ac.id/perpus/main/item/17880
11	Implementasi Sistem Informasi Inventaris Barang Pada Dinas Koperasi Dan UKM Provinsi Kalimantan Tengah Berbasis Web + PDF	E2257401021	2025	Sandi Putra Pratama	MI	https://perpus.stmikplk.ac.id/perpus/main/item/17879
12	Aplikasi Web Profil Desa Sebagai Media Informasi (Studi Kasus Desa Tabak Kanilan) + PDF	E2257401013	2025	Ciptano Nugroho	MI	https://perpus.stmikplk.ac.id/perpus/main/item/17878
13	Perancangan UI / UX Untuk Layanan Pengaduan Masyarakat Berbasis Mobile (Studi Kasus Desa Daya Manunggal) + PDF	E2257401008	2025	Putri Sindi Karolina	MI	https://perpus.stmikplk.ac.id/perpus/main/item/17877
14	Analisis Dan Perancangan Sistem Informasi Arsip Obat Pada Puskesmas Jabiren Raya	E2257401009	2025	Kevin Herwanda	MI	https://perpus.stmikplk.ac.id/perpus/main/item/17876
15	Aplikasi Pendataan Dokumen Pada Kantor Dinas Pertanian Dan Ketahanan Panagn Kota Palangka Raya Berbasis Web + PDF	E2257401023	2025	Radiatul Ikhsan	MI	https://perpus.stmikplk.ac.id/perpus/main/item/17875
16	Aplikasi Arsip Dokumen Sub. Bagian Umum Dan Kepegawaian Pada Dinas Kependudukan Dan Pencatatan Sipil	E2257401016	2025	Reinhard Pebriatama	MI	https://perpus.stmikplk.ac.id/perpus/main/item/17874

Scraping Results | Sheet1 | Sheet2

Gambar 4.3 Upload File *xlsx* ke *Spreadsheets*

Agar data pada *Google Spreadsheets* dapat diakses oleh sistem *website*, *spreadsheet* dipublikasikan terlebih dahulu melalui fitur *Publish to Web* pada *Google Spreadsheets* dengan format CSV. Proses publishing dilakukan dengan membuka menu File, kemudian memilih *Share*, lalu *Publish to Web*, memilih format CSV, dan mengklik tombol *Publish*.



Gambar 4.4 Upload Data Judul ke Publik

Setelah *spreadsheet* berhasil dipublikasikan, langkah selanjutnya adalah mengambil nilai GID dari setiap halaman sheet yang ada. Nilai GID merupakan identifikasi unik yang dimiliki oleh setiap halaman *sheet* pada *Google Spreadsheets* dan dapat dilihat pada URL halaman sheet yang bersangkutan. Nilai GID inilah yang digunakan sistem website untuk mengidentifikasi dan mengambil data dari masing-masing sheet secara terpisah.

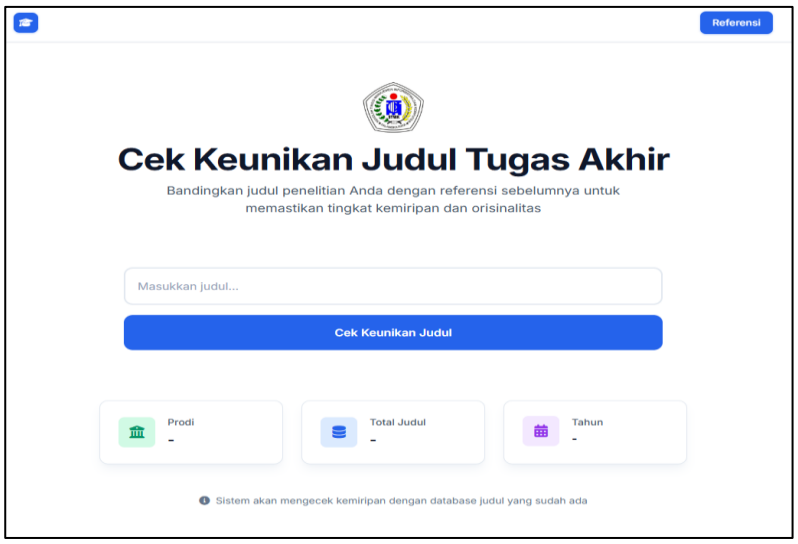


Gambar 4.5 Lokasi GID

b. Sistem *Website*

1) Halaman Utama

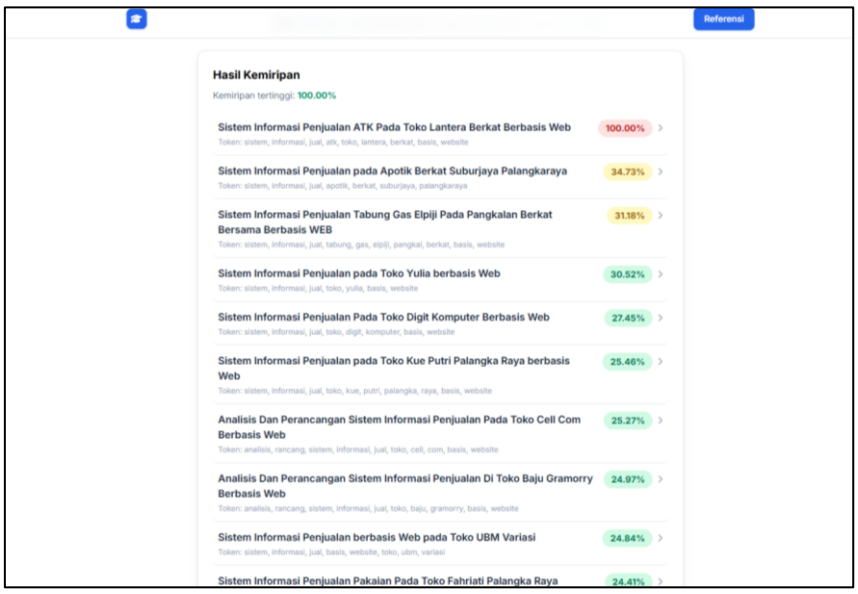
Halaman Utama adalah halaman yang akan pertamakali diakses oleh pengguna yang didalamnya berisikan *navbar* sebuah *text-input* untuk memasukkan judul yang akan dicek serta data statistik yang teredia di database.



Gambar 4.6 Halaman Utama

2) Halaman Hasil

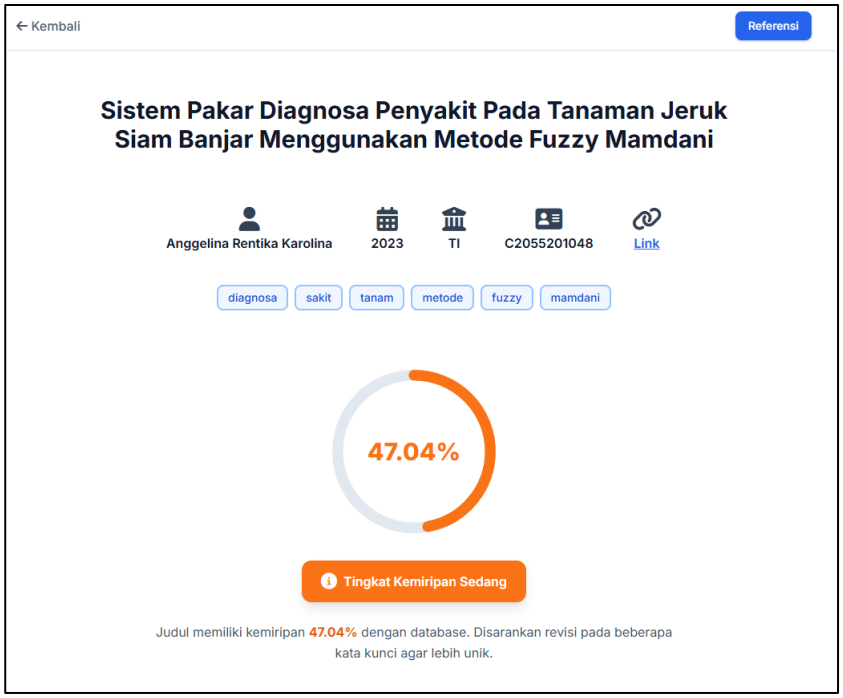
Halaman hasil menampilkan 10 judul paling mirip setelah pengguna melakukan pencarian. Tingkat kemiripan tiap judul direpresentasikan melalui kategori warna: merah (>71%) untuk kemiripan tinggi/signifikan, kuning (31% - 71%) untuk kemiripan sedang, dan hijau (<31%) untuk kemiripan rendah. Pengguna dapat mengeklik setiap hasil untuk dialihkan ke halaman detail.



Gambar 4.7 Halaman Hasil

3) Halaman Detail

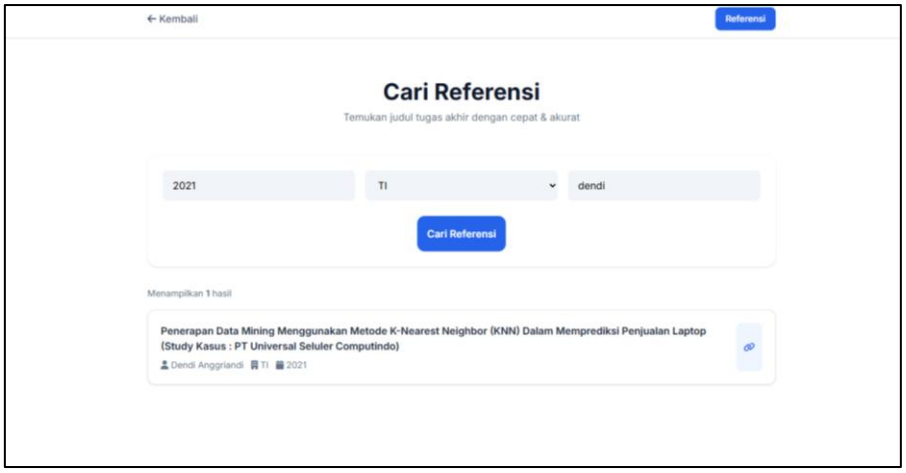
Halaman detail dapat diakses dengan mengeklik salah satu hasil pencarian. Halaman ini menyajikan informasi lengkap tugas akhir berupa judul, penulis, NIM, jurusan, tahun, dan tautan ke perpustakaan STMIK Palangkaraya. Selain itu, ditampilkan pula persentase dan keterangan tingkat kemiripan, serta visualisasi kata-kata yang sama antara judul input dengan judul pembanding.



Gambar 4.8 Halaman Detail

4) Halaman Referensi

Halaman referensi adalah halaman yang digunakan untuk mencari referensi berdasarkan tahun tugas akhir, jurusan dan nama penulis.



Gambar 4.9 Halaman Referensi

c. Sistem *Scraping*

Proses *scraping* berhasil mengambil sebanyak 2340 data judul tugas akhir dari *website* Perpustakaan STMIK Palangkaraya. Data yang berhasil dikumpulkan mencakup atribut judul, nama penulis, NIM, jurusan, tahun, dan tautan file dokumen. Seluruh data tersebut disimpan secara otomatis ke dalam file *Excel* (.xlsx) yang kemudian digunakan sebagai database utama pada sistem website pendeteksi kemiripan judul. Sistem scraping juga dilengkapi dengan mekanisme pengecekan duplikasi, sehingga data yang sudah ada sebelumnya tidak akan tersimpan ganda ke dalam database.

4.1.2 Implementasi Sistem

a. *Preprocessing Text*

Sistem melakukan tahapan preprocessing pada seluruh data judul tugas akhir sebelum proses perhitungan kemiripan dilakukan. Tahapan *preprocessing* yang diterapkan meliputi *case folding, tokenizing, stopword removal*, dan *stemming* yang dijalankan secara berurutan sebagai satu kesatuan proses. Sistem tidak menampilkan hasil antara dari setiap tahapan secara terpisah, melainkan langsung menampilkan hasil akhir berupa token kata yang telah melalui seluruh proses *preprocessing* sekaligus beserta persentase kemiripannya.. Proses *preprocessing* yang diterapkan meliputi beberapa tahapan berikut:

1) Proses mengubah seluruh huruf menjadi huruf kecil dan menghapus karakter selain huruf dan angka.

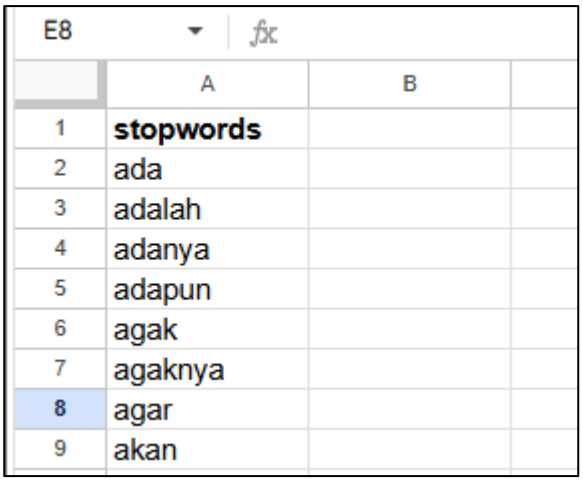
```
const tokens = judul
      .replace(/[^a-z0-9\s]/gi,  "")
      .toLowerCase()
```

2) Proses memecah kalimat menjadi perkata

```
const tokens = judul
      .trim()
```

3) Proses menghapus kata yang dianggap kurang penting

```
const filtered = tokens.filter(word => !stopwords.includes(word));
```



Gambar 4.10 *Stopwords Database*

Daftar *stopword* pada sistem tidak ditulis secara langsung di dalam kode program, melainkan disimpan pada *Google Spreadsheet* agar lebih mudah diperbarui dan dikelola. Sistem akan mengambil data stopword tersebut melalui URL publik dalam format CSV, kemudian menyimpannya ke dalam variabel *stopwords* untuk digunakan pada proses penyaringan kata.

4) Proses mengubah kata menjadi bentuk dasar

```
return stemmer.stem(word);
if (kamusKoreksi[word]) return kamusKoreksi[word];
```



	A	B
	kata_asli	kata_stem
	berbasis	basis
	pengelolaan	kelola
	web	website
	Palangka raya	Palangkaraya
	Perancangan	rancang
	rancang	rancang
	penilaian	nilai

Gambar 4.11 Kamus tambahan Stemming

Pada sistem ini terdapat kamus koreksi kata yang digunakan untuk menyesuaikan beberapa kata tertentu ke bentuk kata sederhana yang sebelumnya terdapat ketidaktepatan pada *library* sastrawi.js. Kamus koreksi tersebut disimpan pada *Google Spreadsheet* yang sama dengan *database*.

Untuk memverifikasi bahwa seluruh tahapan preprocessing berjalan sesuai rancangan, dilakukan pengujian dengan memasukkan data contoh yang sama dengan perhitungan manual Bab III. Hasil akhir token yang ditampilkan sistem dibandingkan dengan hasil akhir stemming pada Bab III ditunjukkan pada Tabel 4.1

Tabel 4.1 Hasil Akhir Token Manual dan Sistem

Judul	Hasil Akhir Token manual	Hasil Akhir Token Sistem
Q	[terap, metode, fuzzy, mamdani, diagnosa, sakit, pata, tanam, peria]	[terap, metode, fuzzy, mamdani, diagnosa, sakit, pata, tanam, peria]
D1	[sistem, pakar, diagnosa, sakit, tanam, jeruk, siam, banjar, metode, fuzzy, mamdani]	[sistem, pakar, diagnosa, sakit, tanam, jeruk, siam, banjar, metode, fuzzy, mamdani]
D2	[sistem, pakar, diagnosa, hama, sakit, tanam, sukun, metode, certainty, factor]	[sistem, pakar, diagnosa, hama, sakit, tanam, sukun, metode, certainty, factor]

b. Implementasi TF-IDF

Proses pembobotan TF-IDF diimplementasikan melalui tiga tahapan, yaitu perhitungan *Term Frequency* (TF), perhitungan *Inverse Document Frequency* (IDF), dan pembentukan vektor TF-IDF. Berbeda dengan perhitungan manual pada Bab III yang menggunakan 3 dokumen contoh untuk keperluan ilustrasi, pada implementasi sistem yang sesungguhnya nilai IDF dihitung terhadap keseluruhan 2340 judul yang tersimpan dalam *database*. Hal ini menyebabkan nilai IDF dan TF-IDF yang dihasilkan sistem akan berbeda secara numerik dengan perhitungan manual Bab III, namun prinsip dan rumus yang digunakan tetap sama.

1) Fungsi *Term Frequency* (TF)

```
function hitungTF(tokens) {
        const tf = {};
        tokens.forEach(token => {
            tf[token] = (tf[token] || 0)+1;
        });

        Object.keys(tf).forEach(k => {
            tf[k] = tf[k] / tokens.length;
        });
        return tf;
}
```

2) Fungsi *Inverse Document Frequency* (IDF)

```
function hitungIDF(semuaTokens) {
        const N = semuaTokens.length;
        const df = {};

        semuaTokens.forEach(tokens => {
            const unik = new Set(tokens);

            unik.forEach(token => {
                df[token] = (df[token] || 0) + 1;
```

```
            });
        });

        const idf = {};
        Object.keys(df).forEach(token => {
            idf[token] = Math.log((N + 1) / (df[token] + 1)) + 1;
        });
        return idf;
}
```

Tahap ini mengukur seberapa penting suatu kata di antara seluruh judul yang ada di *database*. Kata yang muncul di banyak judul dianggap kurang penting dan diberi bobot rendah, sedangkan kata yang jarang muncul dianggap lebih spesifik dan diberi bobot tinggi.

3) Fungsi Pembentukan Vektor TF-IDF

```
function buatVektor(tokens, idf) {
        const tf = hitungTF(tokens);
        const vektor = {};

        Object.keys(tf).forEach(token => {
            vektor[token] = tf[token] * (idf[token] ||
    Math.log((semuaData.length + 1) / (1 + 1)) + 1);
        });
        return vektor;
}
```

Tahap ini menggabungkan nilai TF dan IDF untuk setiap kata dengan cara mengalikan keduanya, sehingga menghasilkan satu nilai bobot untuk setiap kata dalam sebuah judul. Nilai bobot inilah yang merepresentasikan seberapa penting suatu kata dalam konteks judul tersebut dibandingkan dengan seluruh judul di *database*.

c. Implementasi Cosine Similarity

```
function cosineSimilarity(vektorA, vektorB) {
  const semuaKey = new Set([...Object.keys(vektorA),
...Object.keys(vektorB)]);

  let dotProduct = 0;
```

```
    let magA = 0;
    let magB = 0;

    semuaKey.forEach(key => {
        const a = vektorA[key] || 0;
        const b = vektorB[key] || 0;

        dotProduct += a * b;
        magA += a * a;
        magB += b * b;
    });

    if (magA === 0 || magB === 0) return 0;

    return dotProduct / (Math.sqrt(magA) * Math.sqrt(magB));
}
```

Hasil *Cosine Similarity* yang ditampilkan sistem merupakan hasil perhitungan terhadap keseluruhan 2340 judul dalam *database*. Hasil yang ditampilkan sistem ditunjukkan pada Tabel 4.2

Tabel 4.2 Hasil Cosine Similarity pada Sistem

Judul	Cosine Similarity Sistem	Kategori
Sistem Pakar Diagnosa Penyakit Pada Tanaman Jeruk Siam Banjar Menggunakan Metode Fuzzy Mamdani	48.75%	Sedang
Sistem Pakar Diagnosa Hama Dan Penyakit Pada Tanaman Sukun Menggunakan Metode Certainty Factor	30.18%	Rendah

d. Verifikasi Perhitungan Manual Terhadap Sistem

Untuk memastikan bahwa implementasi algoritma pada sistem berjalan sesuai dengan prinsip matematis yang telah dirancang, dilakukan verifikasi dengan membandingkan hasil perhitungan manual dengan keluaran sistem menggunakan data contoh yang sama.

Hasil perbandingan antara perhitungan manual Bab III dengan output sistem ditunjukkan pada Tabel 4.3

Tabel 4.3 Perbandingan Cosine Similarity Sistem dan Manual

Dokumen	Perhitungan Manual	Kategori Manual	Perhitungan Sistem	Kategori Sistem	Keterangan
D1	46.3%	Sedang	48.75%	Sedang	Kategori Sesuai
D2	27.9%	Rendah	30.18%	Rendah	Kategori Sesuai

4.2 Pengujian

4.2.1 Uji *Blackbox*

Uji *Black Box* dilakukan untuk memastikan bahwa sistem berjalan sesuai dengan fungsi yang diharapkan tanpa melihat proses internal di dalamnya. Pengujian *black box* dilakukan oleh 3 orang penguji yang terdiri dari dosen STMIK Palangkaraya. Pengujian ini dibagi menjadi dua bagian, yaitu pengujian terhadap sistem utama deteksi kemiripan judul dan pengujian terhadap aplikasi scraping.

Tabel 4.4 Uji blackbox pada sistem website

Halaman	**Output yang diharapkan**	**Hasil**
Halaman Utama	Sistem Menampilkan Halaman Utama	[✓] Sesuai [ ] Tidak Sesuai
	Sistem dapat memproses input judul dan menampilkan hasil kemiripan	[✓] Sesuai [ ] Tidak Sesuai
Halaman Hasil	Sistem menampilkan 10 judul beserta persentase kemiripan	[✓] Sesuai [ ] Tidak Sesuai
Halaman Detail	Sistem menampilkan detail data judul yang dipilih	[✓] Sesuai [ ] Tidak Sesuai

Halaman Referensi	Sistem menampilkan data referensi dengan baik	[✓] Sesuai [ ] Tidak Sesuai

Tabel 4.5 Uji *Blackbox* pada aplikasi *scraping*

Skenario Uji	Hasil yang diharapkan	Status
Koneksi ke *Website* Perpustakaan	Sistem berhasil mengakses *website* perpustakaan STMIK Palangkaraya	[✓] Sesuai [ ] Tidak Sesuai
Pengambilan Data	Sistem berhasil mengambil data judul tugas akhir	[✓] Sesuai [ ] Tidak Sesuai
Penyaringan Data Duplikat	Sistem tidak menyimpan data yang sudah ada sebelumnya	[✓] Sesuai [ ] Tidak Sesuai
Penyimpanan data *Excel*	Sistem berhasil menyimpan data hasil *scraping* ke file .xlsx	[✓] Sesuai [ ] Tidak Sesuai
Program *executable* (.exe)	Program dapat dijalankan tanpa instalasi *Python*	[✓] Sesuai [ ] Tidak Sesuai

4.2.2 Uji Validasi Algoritma

a. Penilaian Pakar

Uji validasi dilakukan untuk mengukur kesesuaian label kemiripan sistem dengan penilaian pakar. Pengujian ini melibatkan 3 orang dosen untuk menilai 30 pasangan judul skripsi ke dalam kategori Rendah, Sedang, dan Tinggi melalui kuesioner, di mana label akhir pakar ditentukan berdasarkan nilai modus. Label modus pakar kemudian dibandingkan dengan hasil *Cosine Similarity* sistem berdasarkan *threshold* yang ditetapkan: Rendah (<30%), Sedang (31% - 71%), dan Tinggi (>71%). Hasil perbandingan kedua label disajikan pada tabel berikut.

Tabel 4.6 Hasil Penilaian Kemiripan oleh pakar

No	Pasangan Judul	Hasil Manual	Hasil Sistem	Label Pakar
1.	judul I: Sistem Penentuan Kelayakan Kenaikan Jabatan Pada Karyawan Pizza Hut Palangka Raya Menggunakan Metode Profile Matching judul II: Prototype Alat Pakan Ikan Otomatis Menggunakan Aplikasi Blynk Berbasis Internet Of Things	0,00	0	Rendah
2.	Judul I : Penerapan Metode Topsis Dalam Pemilihan Jenis Pupuk Kebun Sawit Pada Lahan Desa Kenyala Judul II : Analisis Dan Perancangan Sistem Informasi Layanan Imunisasi Pada Posyandu Desa Palanggan Berbasis web	5,58	0	Rendah
3.	Judul I: Sistem Informasi Pengaduan Masyarakat Pada Dinas Pekerjaan Umum Kabupaten Gunung Mas Berbasis Web Judul II : Aplikasi Perpustakaan Pada Kantor Pengadilan Tata Usaha Negara Kota Palangkaraya Berbasis Web	11,07	5,68	Rendah
4.	Judul I : Penerapan Arsitektur Efficientnet-B1 Untuk Klasifikasi Tanaman Hias Tropis Berbasis CNN Judul II : Sistem Pakar Deteksi Dini Gizi Buruk Pada Balita Usia 12-59 Bulan Menggunakan Metode Forward Chaining	0,00	0	Rendah
5.	Judul I : Analisis Dan Perancangan Sistem Informasi Buku Tamu Pada Dinas Energi Dan Sumber Daya Mineral Provinsi Kalimantan Tengah Judul II : Analisis Dan Perancangan Sistem Informasi Pengarsipan Pada Dinas Perumahan Rakyat, Kawasan Permukiman Dan Pertanahan Kota Palangka Raya	25,81	17,6	Rendah

6.	Judul I : Penerapan Rule-Based Pada Sistem Peminjaman Sarana Dan Prasarana DISPORA Kalimantan Tengah Judul II : Penerapan Augmented Reality (AR) Sebagai Pengenalan Tempat Wisata Di Kota Palangkaraya	6,88	8,47	Rendah
7	Judul I : Aplikasi Kehadiran Mahasiswa menggunakan Barcode berbasis Android Studi Kasus STMIK Palangka Raya Judul II : Sistem Pendukung Keputusan Seleksi Mahasiswa Berprestasi pada STMIK Palangka Raya menggunkan Metode Simple Additive Weighting (SAW)	21,96	21,01	Rendah
8	Judul I : Perancangan Desain Antar Muka Website E-Commerce Pertukangan Di Palangkaraya Judul II : Klasifikasi Citra Hewan Khas Dayak Dengan Menggunakan CNN	0,00	0	Rendah
9	Judul I : Penerapan Metode Naïve Bayes Untuk Mendiagnosis Penyakit Dan Hama Pada Tanaman Kangkung Judul II : Penerapan Data Mining Dalam Menentukan Kinerja Karyawan Terbaik Berbasis Algoritma C45 Pada Balai Karantina Hewan, Ikan Dan Tumbuhan Kalimantan Tengah	4,95	0	Rendah
10	Judul I : Aplikasi Pengenalan Huruf, Angka, dan Warna dengan Bahasa Inggris Untuk Anak Usia 3-6 Tahun Berbasis Android Judul II : Aplikasi Jadwal Mata Kuliah pada STMIK Palangka Raya berbasis Android	19,06	11,84	Rendah
11	Judul I : Sistem Pakar Diagnosis Penyakit Orang Utan Menggunakan Metode Certainty Factor Judul II : Aplikasi Sistem Pakar Diagnosis Penyakit Hepatitis berbasis Web menggunakan Metode Certainty Factor	58,53	64,31	Tinggi

12	Judul I : Analisis Dan Perancangan Sistem Informasi Inventaris Barang Berbasis Website Di Kantor Perpustakaan Dan Kearsipan Daerah Palangka Raya Judul II : Analisis Dan Perancangan Sistem Informasi Inventaris Barang Pada STMIK Palangkaraya	57,32	65,06	Tinggi
13	Judul I : Analisis Dan Perancangan Sistem Informasi Koperasi Simpan Pinjam Sejahtera STMIK Palangkaraya Berbasis WEB Judul II : Dan Perancangan Sistem Koperasi Simpan Pinjam Pada SMK YPSEI Palangka Raya Berbasis Web	61,69	67,52	Tinggi
14	Judul I : Rancang Bangun Sistem Informasi Geografis Pemetaan Tempat Pembuangan Sampah Sementara Di Daerah Kota PalangkaRaya Judul II : Analisis Dan Perancangan Sistem Informasi Geografis Berbasis web Utntuk Pemetaan Lokasi Salon Kecantikan Di Kota Palangkaraya	40,80	40,73	Sedang
15	Judul I : Diagnosis Penyakit Dan Hama Pada Buah Mengkudu Dengan Metode Forward Chaining Berbasis Web Judul II : Sistem Pakar Identifikasi Penyakit Pada Buah Nanas Menggunakan Metode Certainty Factor Dan Forward Chaining Berbasis Web	46,39	49,59	Sedang
16	Judul I : Prediksi Analisis Sentimen Unggahan Instragram DITSAMAPTA POLDA Kalimantan Tengah Menggunakan Naïve Bayes Judul II : Analisis Sentimen Menggunakan Naïve Bayes Pada Berita Kepolisian POLDA Kalimantan Tengah	61,69	56,37	Tinggi
17	Judul I : Aplikasi Penjadwalan Laboratorium STMIK Palangkaraya Menggunakan Metode Algoritma Genetika Judul II : Penerapan Metode Algoritma Genetika Untuk Penjadwalan Tugas	42,67	60,69	Sedang

	Akhir Program Studi Teknik Informatika STMIK Palangkaraya			
18	Judul I : Game Edukasi Augmented Reality Pengenalan Hewan Berdasarkan Habitat Judul II : Penerapan Augmented Reality (AR) Sebagai Pengenalan Tempat Wisata Di Kota Palangkaraya	24,43	32,92	Sedang
19	Judul I : Rancang Bangun Pemetaan Objek Wisata Di Kabupaten Kotawaringin Barat Berbasis Gis Judul II : Rancang Bangun Sistem Informasi Geografis Pemetaan Lokasi Objek Wisata Di Kabupaten Gunung Mas Menggunakan Quantum GIS	46,39	57,23	Tinggi
20	Judul I : Sistem Pendukung Keputusan Penentuan Karyawan Terbaik Pada STMIK Palangkaraya Menggunakan Metode Simple Additive Weighting (SAW) Judul II : Sistem Pendukung Keputusan Penentuan Dosen Pembimbing Tugas Akhir Terbaik Prodi Teknik Informatika STMIK Palangkaraya Menggunakan Metode Simple Additive Weighting	62,65	64,54	Tinggi
21	Judul I : Sistem Informasi Skripsi dan Tugas Akhir STMIK Palangka Raya Berbasis Web Judul II : Aplikasi E-Library Tugas Akhir dan Skripsi Pada STMIK Palangka Raya Berbasis Web	67,40	73,81	Tinggi
22	Judul I : Analisis Dan Perancangan Sistem Informasi Pengarsipan Pada Dinas Perumahan Rakyat, Kawasan Permukiman Dan Pertanahan Kota Palangka Raya Judul II : Sistem Informasi Pendataan Pertanahan Di Dinas Perumahan Rakyat, Kawasan Permukiman Dan Pertanahan Kota Palangkaraya Berbasis Website	66,63	83,65	Tinggi
23	Judul I : Analisis Dan Perancangan Sistem Informasi Pengajuan Program Bantuan Pasang Baru Listrik Pada Dinas Energi Dan Sumber Daya	63,37	72,21	Tinggi

	Mineral Provinsi Kalimantan Tengah Judul II : Analisis Dan Perancangan Pengajuan Kegiatan Kedinasan Pada Dinas Energi Dan Sumber Daya Mineral Provinsi Kalimantan Tengah			
24	Judul I : Aplikasi Kebudayaan Dayak Ngaju Di Kalimantan Tengah Berbasis Mobile Judul II : Rancang Bangun Aplikasi Budaya Dayak Ngaju Kalimantan Tengah Berbasis Web Mobile	78,21	90,16	Tinggi
25	Judul I : Analisis Dan Perancangan Sistem Informasi Unit Kegiatan Mahasiswa (UKM) Pada STMIK Palangkaraya Judul II : Analisis Dan Perancangan Sistem Informasi Unit Kegiatan Mahasiswa (UKM) KSR STMIK Palangkaraya Berbasis WEB	81,44	84,14	Tinggi
26	Judul I : Rancang Bangun Aplikasi Pengenalan Alat Musik Kalimantan Tengah Berbasis Web Mobile Judul II : Aplikasi Pengenalan dan Alat Musik Tradisional Kalimantan Tengah	56,88	70,83	Tinggi
27	Judul I : Analisis Dan Perancangan Sistem Informasi Pengajuan Cuti Pegawai Berbasis Web Pada Kantor Badan Pengelola Pajak Dan Retribusi Daerah Kota Palangka Raya Judul II : Analisis dan Perancangan Sistem Informasi Manajemen Kearsipan pada Badan Pengelola Pajak dan Retribusi Daerah Kota Palangka Raya berbasis Web	72,23	75,06	Sedang
28	Judul I : Aplikasi Panduan Pramuka Berbasis Mobile Judul II : Aplikasi Pengenalan Panduan Materi Pramuka berbasis Web Mobile	70,44	76,46	Tinggi
29	Judul I : Sistem Informasi Unit Kegiatan Mahasiswa (UKM) Kewirausahaan STMIK Palangkaraya Berbasis Web Judul II : Analisis Dan Perancangan Sistem Informasi Unit Kegiatan	65,56	79,05	Tinggi

	Mahasiswa (UKM) Pada STMIK Palangkaraya			
30	Judul I : Sistem Informasi Tata Kelola Kepariwisataan pada Dinas Pariwisata dan Ekonomi Kreatif Kota Palangka Raya Judul II : Sistem Informasi Pengelolaan Data Inventaris Barang dan Aset pada Dinas Pariwisata dan Ekonomi Kreatif Kota Palangka Raya	65,08	74,15	Sedang

b. Hasil klasifikasi

Berdasarkan hasil perbandingan antara label sistem dan label pakar pada tabel di atas, selanjutnya dilakukan rekapitulasi ke dalam *confusion matrix* untuk melihat distribusi kesesuaian dan ketidaksesuaian label pada masing-masing kelas. *Confusion matrix* hasil perbandingan label sistem dan label pakar disajikan pada tabel berikut.

Tabel 4.7 *Confussion Matrix*

	Pred: Rendah	Pred: Sedang	Pred: Tinggi
Aktual: Rendah	10	0	0
Aktual: Sedang	0	4	6
Aktual: Tinggi	0	2	8

Berdasarkan nilai *Confusion Matrix* yang telah diperoleh, selanjutnya dihitung nilai *Precision*, *Recall*, dan *F1-Score* untuk mengukur tingkat kinerja sistem dalam mengklasifikasikan kemiripan judul skripsi pada setiap kelas. Hasil perhitungan disajikan pada tabel berikut.

Tabel 4.8 Hasil Kinerja Sistem

Kelas	**Precision**	**Recall**	**F1-Score**
Rendah	1.000	1.000	1.000
Sedang	0.677	0.400	0.500
Tinggi	0.746	0.800	0.677

Nilai akurasi sistem dihitung dari rasio kecocokan klasifikasi sistem dengan penilaian pakar terhadap total data uji. Berdasarkan 30 pasangan judul yang diujikan, sistem berhasil mengklasifikasikan 22 data secara benar, yang terdiri dari 10 pasangan pada kelas Rendah, 4 pasangan pada kelas Sedang, dan 8 pasangan pada kelas Tinggi. Perhitungan akurasi sistem diperoleh sebagai berikut:

$$Akurasi = \frac{22}{30} \times 100\% = 73{,}3\%$$

c. Penyesuaian *Threshold*

Berdasarkan hasil evaluasi pada pengujian sebelumnya, penggunaan *threshold* 71% untuk kelas Tinggi menghasilkan metrik performa yang belum optimal secara aktual dinilai mirip oleh pakar. Untuk mengatasi batasan tersebut dan mencari titik kalibrasi yang lebih baik, perlu dilakukan penyesuaian parameter. Pada bagian ini, akan dipaparkan hasil evaluasi ulang performa klasifikasi sistem ketika batas threshold kemiripan kelas Tinggi diturunkan menjadi 60%, yang bertujuan untuk meningkatkan akurasi keseluruhan agar lebih representatif terhadap penilaian pakar menghasilkan prediksi sebagai berikut:

Tabel 4.9 Confusion Matrix Setelah Penyesuaian

	Pred: Rendah	Pred: Sedang	Pred: Tinggi
Aktual: Rendah	10	0	0
Aktual: Sedang	0	3	3

Aktual: Tinggi	0	2	12

Perubahan threshold dari 71% menjadi 60% menyebabkan beberapa data yang sebelumnya dikategorikan sebagai "Sedang" berpindah ke kategori "Tinggi". Hal ini mengakibatkan distribusi hasil klasifikasi menjadi tidak seimbang. Namun, kondisi tersebut merupakan konsekuensi dari penyesuaian batas klasifikasi dan tidak memengaruhi validitas perhitungan metrik evaluasi seperti *accuracy, precision, recall,* dan *F1-score*.

Tabel 4.10 Hasil Kinerja Sistem Setelah Penyesuaian

Kelas	**Precision**	**Recall**	**F1-Score**
Rendah	1.000	1.000	1.000
Sedang	0.600	0.500	0.545
Tinggi	0.800	0.857	0.828

Nilai akurasi sistem dihitung dari rasio kecocokan klasifikasi sistem dengan penilaian pakar terhadap total data uji. Berdasarkan 30 pasangan judul yang diujikan, sistem berhasil mengklasifikasikan 25 data secara benar, yang terdiri dari 10 pasangan pada kelas Rendah, 3 pasangan pada kelas Sedang, dan 12 pasangan pada kelas Tinggi. Perhitungan akurasi sistem diperoleh sebagai berikut:

$$Akurasi = \frac{25}{30} \times 100\% = 83{,}3\%$$

d. Perbandingan *Threshold*

Tabel 4.11 Perbandingan Kinerja Sistem sebelum dan Sesudah Penyesuaian

Kelas	Metriks	Threshold 71%	Treshold 60%
Rendah	Precision	1.000	1.000
	Recall	1.000	1.000

	F1-Score	1.000	1.000
Sedang	Precision	0.677	0.600
	Recall	0.400	0.500
	F1-Score	0.500	0.545
Tinggi	Precision	0.746	0.800
	Recall	0.800	0.857
	F1-Score	0.677	0.828
Akurasi		73,3%	83,3%

4.3 Pembahasan

Berdasarkan hasil pengujian *black box* yang telah dilakukan terhadap *website* dan aplikasi *scraping*, seluruh skenario pengujian menunjukkan hasil yang sesuai dengan yang diharapkan. Aplikasi scraping juga berhasil mengumpulkan 2340 data judul skripsi dari *website* perpustakaan STMIK Palangkaraya dan tersimpan dengan baik ke dalam *database spreadsheets*. Hal ini menunjukkan bahwa seluruh komponen sistem telah berfungsi dengan baik dan sesuai dengan kebutuhan pengguna.

Berdasarkan hasil uji validasi label menggunakan *threshold* awal sebesar 71%, sistem memperoleh nilai akurasi sebesar 73,3% dari 30 pasangan judul yang diujikan. Hasil tersebut menunjukkan bahwa sebanyak 22 pasangan judul berhasil diklasifikasikan sesuai dengan penilaian pakar, sedangkan 8 pasangan judul lainnya mengalami ketidaksesuaian label antara sistem dan pakar. Pada kelas Rendah, sistem menunjukkan kinerja yang sangat baik dengan nilai *Precision, Recall*, dan *F1-Score* masing-masing sebesar 1,000, yang mengindikasikan bahwa sistem mampu mengidentifikasi seluruh pasangan judul berkemiripan rendah secara tepat sesuai dengan penilaian pakar. Pada kelas Tinggi, sistem memperoleh nilai *Recall*

sebesar 0,800, yang menunjukkan bahwa sebagian besar pasangan judul dengan tingkat kemiripan tinggi telah berhasil dikenali oleh sistem. Sementara itu, pada kelas Sedang diperoleh nilai *F1-Score* terendah yaitu sebesar 0,500, di mana 6 dari 10 pasangan judul yang menurut pakar termasuk dalam kategori Sedang justru diklasifikasikan sebagai Tinggi oleh sistem. Kondisi ini mengindikasikan bahwa *threshold* pemisah antara kelas Sedang dan Tinggi yang ditetapkan pada nilai 71% kurang optimal untuk konteks kemiripan judul skripsi.

Untuk mengatasi permasalahan tersebut, dilakukan pengujian lanjutan dengan melakukan penyesuaian *threshold* kategori Tinggi dari 71% menjadi 60%, berdasarkan hasil kuesioner pakar yang menunjukkan bahwa sebagian besar dosen mulai menganggap pasangan judul memiliki kemiripan tinggi pada rentang 60% ke atas. Hasil pengujian menunjukkan bahwa perubahan *threshold* tersebut mampu meningkatkan performa sistem. Berdasarkan evaluasi menggunakan *confusion matrix*, sistem memperoleh nilai akurasi sebesar 83,3%, dengan 25 dari 30 pasangan judul berhasil diklasifikasikan sesuai dengan penilaian pakar. Pada skenario threshold 60%, kelas Rendah tetap menunjukkan kinerja sempurna dengan nilai *Precision, Recall*, dan *F1-Score* sebesar 1,000. Selain itu, kelas Tinggi mengalami peningkatan performa dengan nilai Precision sebesar 0,800, *Recall* sebesar 0,857, dan *F1-Score* sebesar 0,828. Meskipun kelas Sedang masih menjadi kelas dengan performa terendah, nilai Recall meningkat dari 0,400 menjadi 0,500, sedangkan nilai *F1-Score* meningkat dari 0,500 menjadi 0,545.

Secara keseluruhan, hasil penelitian menunjukkan bahwa sistem telah mampu mengklasifikasikan kemiripan judul skripsi dengan cukup baik, khususnya pada kelas Rendah dan Tinggi. Perbandingan antara *threshold* 71% dan 60% menunjukkan bahwa penggunaan threshold 60% lebih sesuai dengan persepsi pakar dalam menilai kemiripan judul skripsi. Penyesuaian *threshold* tersebut terbukti dapat meningkatkan akurasi sistem dari 73,3% menjadi 83,3%, serta meningkatkan nilai *Recall* dan *F1-Score* pada sebagian besar kelas. Oleh karena itu, *threshold* 60% dapat direkomendasikan sebagai batas klasifikasi kemiripan tinggi pada sistem deteksi kemiripan judul skripsi yang dikembangkan dalam penelitian ini.

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan penelitian yang telah dilakukan, penulis dapat meyimpulkan beberapa hal yaitu:

a. Penelitian ini berhasil membangun sebuah aplikasi sistem berbasis *website* untuk mendeteksi kemiripan judul tugas akhir menggunakan metode TF-IDF dan *Cosine Similarity*. Dalam implementasinya, data pada sistem ini telah terintegrasi secara dinamis dengan *website* perpustakaan STMIK Palangkaraya melalui aplikasi *web scraping* yang telah dikembangkan.
b. Semua fitur yang ada di sistem ini dapat digunakan dengan baik sesuai dengan fungsinya setelah dilakukan pengujian *blackbox*.
c. Berdasarkan pengujian performa klasifikasi menggunakan *threshold* awal sebesar 71%, sistem memperoleh nilai akurasi sebesar 73,3% dengan 22 dari 30 pasangan judul berhasil diklasifikasikan sesuai dengan penilaian pakar. Pada skenario ini, Kelas Rendah menunjukkan performa sempurna dengan nilai *Precision, Recall*, dan *F1-Score* sebesar 1,000. Pada Kelas Tinggi diperoleh nilai *Recall* sebesar 0,800 dengan *F1-Score* sebesar 0,677, sedangkan Kelas Sedang menghasilkan nilai *F1-Score* terendah yaitu sebesar 0,500 (*Precision* 0,677 dan *Recall* 0,400).
d. Pengujian lanjutan dengan melakukan penyesuaian *threshold* kategori Tinggi dari 71% menjadi 60% menunjukkan peningkatan performa sistem. Pada *threshold* 60%, sistem memperoleh nilai akurasi sebesar 83,3%, dengan 25 dari

30 pasangan judul berhasil diklasifikasikan sesuai dengan penilaian pakar. Pada skenario ini, Kelas Rendah tetap menunjukkan performa sempurna dengan nilai *Precision, Recall*, dan *F1-Score* sebesar 1,000. Kelas Tinggi mengalami peningkatan performa dengan nilai *Precision* sebesar 0,800, *Recall* sebesar 0,857, dan *F1-Score* sebesar 0,828. Sementara itu, Kelas Sedang memperoleh nilai *Precision* sebesar 0,600, *Recall* sebesar 0,500, dan *F1-Score* sebesar 0,545.

5.2 Saran

Berdasarkan hasil penelitian, pengujian, serta keterbatasan yang ditemukan pada sistem deteksi kemiripan judul tugas akhir ini, maka saran yang dapat diberikan untuk pengembangan penelitian selanjutnya adalah sebagai berikut:

a. Melakukan kajian lebih lanjut terhadap batas persentase (*threshold*) klasifikasi kemiripan dengan melibatkan jumlah validator dan data uji yang lebih banyak.
b. Melakukan penambahan kosakata pada *database* kamus sinonim yang sudah ada di dalam sistem. Hal ini penting untuk melengkapi variasi kata standar dan tidak standar yang maknanya sama namun belum terdefinisi di dalam sistem (seperti kata "diagnosa" dengan "diagnosis", atau "metode" dengan "metoda"), sehingga kalkulasi pembobotan kata oleh algoritma menjadi jauh lebih akurat.
c. Menerapkan mekanisme otomatisasi penjadwalan agar aplikasi *web scraping* dapat berjalan otomatis di latar belakang untuk memperbarui data ke *database spreadsheet* tanpa perlu intervensi tindakan klik dari pengguna lagi.

DAFTAR PUSTAKA

Albab, M.U., P., Y.K., Fawaiq, M.N., 2023. Optimization of the Stemming Technique on Text Preprocessing President 3 Periods Topic. Jurnal Transformatika 20, 1–12. https://doi.org/10.26623/transformatika.v20i2.5374

Amanah Amnun Zulfa, Tatang Ibrahim, O.A., 2025. Peran Sistem Informasi Akademik Berbasis Web Dalam Upaya Meningkatkan Efektivitas Dan Efisiensi Pengelolaan Akademik Di Perguruan Tinggi. Jurnal Tahsinia 6, 115–134.

Dikilitaş, Y., Çakal, Ç., Okumuş, A.C., Yalçın, H.N., Yıldırım, E., Ulusoy, Ö.F., Macit, B., Kırkaya, A.E., Yalçın, Ö., Erdoğmuş, E., Sayar, A., 2024. Performance Analysis for Web Scraping Tools: Case Studies on Beautifulsoup, Scrapy, Htmlunit and Jsoup. Lecture Notes in Networks and Systems 960, 471–480. https://doi.org/10.1007/978-3-031-56728-5_39

Eko, H., Adi, U., Nasiri, A., Informatika, M.T., Framework, M., 2023. Studi Dan Analisis Algoritmakemiripan Cosine Similarity Untuk Mengidentifikasikesamaan Judul Skripsi. PROSIDING SEMINAR ILMIAH SISTEM INFORMASI DAN TEKNOLOGI INFORMASI XII, 1–9.

Fandopa, J.A., Santoso, N., 2022. Pengembangan Sistem Informasi Manajemen Percetakan pada Gajayana Digital Printing Kota Malang berbasis Website. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer 6, 5371–5379.

Halim, J., Lasut, D., 2024. Document Plagiarism Detection Application Using Web-Based TF-IDF and Cosine Similarity Methods. bit-Tech 7, 202–213. https://doi.org/10.32877/bt.v7i2.1697

Hendrawan, E., Meisel, M., Sari, D.N., 2023. Analysis And Implementation Of Computer Network Systems Using Software Draw.io. Asia Information System Journal 2, 9–15.

Kurniawan, A.A., Mustikasari, M., 2021. Implementasi Deep Learning Menggunakan Metode CNN dan LSTM untuk Menentukan Berita Palsu dalam Bahasa Indonesia. Jurnal Informatika Universitas Pamulang 5, 544. https://doi.org/10.32493/informatika.v5i4.6760

Kurniawan, S., Wiranata, W., Kusnan, K., Ma'muriyah, N., Ting, V.V., 2023. Pemanfaatan Komputasi Awan (Cloud Computing) Pada Bidang Pendidikan. Journal of Information System and Technology 4, 403–405. https://doi.org/10.37253/joint.v4i2.6243

Lumbansiantar, S., Dwiasnati, S., Fatonah, N.S., 2023. Penerapan Metode Cosine Similarity Dalam Mendeteksi Plagiarisme Pada Jurnal. Format : Jurnal Ilmiah Teknik Informatika 12, 142. https://doi.org/10.22441/format.2023.v12.i2.007

Meidelfi, D., Rahmayuni, I., Hidayat, T., Chandra, D., 2021. TF-IDF Implementation for Similarity Checker on The Final Project Title 3, 40–52.

Muhyidin, M.A., Sulhan, M.A., Sevtiana, A., 2020. Perancangan Ui/Ux Aplikasi My Cic Layanan Informasi Akademik Mahasiswa Menggunakan Aplikasi Figma. Jurnal Digit 10, 208. https://doi.org/10.51920/jd.v10i2.171

Nasrullah, A.H., 2024. Integrasi Tf-Idf Dan Algoritma Cosine Similarity Untuk Deteksi Tingkat Kemiripan Judul Penelitian (Studi Kasus Mahasiswa Fakultas Ilmu Komputer UNISAN Gorontalo) 3, 113–118.

Nilawati, Husaini, J.S., 2024. Penggunaan Metode Cosine Similarity Dan Tf-idf Untuk Klasifikasi Judul Seminar Proposal Pada Fakultas Teknik Universitas Jabal Ghafur. Sagita Academia Journal 2, 72–79.

Purwaningrum, S., Susanto, A., Prabowo, A.S., 2025. Kombinasi Algoritma TF-IDF dan Weighted Dice Similarity Untuk Pengukuran Kemiripan Judul Tugas Akhir 16, 328–337. https://doi.org/10.35970/infotekmesin.v16i2.2812

Riyani, A., Zidny, M., Burhanuddin, A., 2019. Penerapan Cosine Similarity dan Pembobotan TF-IDF untuk Mendeteksi Kemiripan Dokumen 2, 23–27.

Salman, 2025. Sistem Pendeteksi Kemiripan Judul Skripsi Berbasis Web Menggunakan Nlp Dan Word Embeddings. Jurnal Pendidikan Indonesia:Teori, Penelitian dan Inovasi 5. https://doi.org/10.59818/jpi.v5i5.1964

Trilaksono, A.R., Husain, T., Doharma, R., 2020. Uji Model Keberhasilan Sistem Informasi: Media Penyimpanan Google Drive. Jurnal Teknologi Sistem Informasi dan Aplikasi 3, 57. https://doi.org/10.32493/jtsi.v3i2.4434

Fitria Nur Hasanah M.Pd, R. S. U. M., 2020. *Rekayasa Perangkat Lunak.* Pertama ed. Sidoarjo: Universitas Muhammadiyah Sidoarjo.

Halim, J., 2024. *Aplikasi Deteksi Plagiarisme Dokumen Dengan Metode Tf-idf Dan Cosine Similarity Berbasis Web,* Tangerang: Universitas Buddhi Dharma.

Suwarya, F. M., 2021. *Dahsyatnya Google Drive.* Indramayu: Guepedia.

L A M P I R A N

SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN KOMPUTER
(STMIK) PALANGKARAYA
Jl. G. Obos No.114 Telp.0536-3224593, 3225515 Fax.0536-3225515 Palangka Raya
email : humas@stmikplk.ac.id - website : www.stmikplk.ac.id

SURAT TUGAS PEMBIMBING TUGAS AKHIR

NO. 285/STMIK-3.C.2/AU/XI/2025

Ketua Program Studi Teknik Informatika Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) Palangkaraya menugaskan nama-nama tersebut di bawah ini :

1. Nama : Dr. Maura. Widyaningsih, S.Kom., M.Cs
   NIK : 197502232009002
   Sebagai Pembimbing I dalam **PEMBUATAN PROGRAM**
2. Nama : Ir. Abdul Hadi, S.T., M.Kom
   NIK : 198505102021101
   Sebagai Pembimbing II dalam **PENULISAN TUGAS AKHIR**

Untuk membimbing Tugas Akhir mahasiswa :

Nama	:	Ibnu Rofik
NIM	:	C2255201109
Program Studi	:	TEKNIK INFORMATIKA (55201)
Tanggal Daftar	:	05/11/2025
Judul Tugas Akhir	:	Penerapan TF-IDF dan Cosine Similarity pada Deteksi Kemiripan Judul Tugas Akhir di STMIK Palangkaraya

Demikian surat ini dibuat agar dapat dipergunakan sebagaimana mestinya, dan dilaksanakan dengan penuh tanggung jawab.

Palangka Raya, 07 November 2025

Ketua Program Studi,

November 2025

Ir. Abdul Hadi, S.T., M.Kom
NIK. 198505102021101

TEMBUSAN :
1. Pembimbing I dan II
2. Mahasiswa yang bersangkutan
3. Arsip Prodi TI

SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN KOMPUTER
(STMIK) PALANGKARAYA
Jl. G. Obos No.114 Telp.0536-3224593, 3225515 Fax.0536-3225515 Palangka Raya
email : humas@stmikplk.ac.id - website : www.stmikplk.ac.id

SURAT TUGAS PENGUJI TUGAS AKHIR

537/STMIK-3.C.2/AU/V/2026

Ketua Program Studi Teknik Informatika Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) Palangkaraya menugaskan kepada nama- nama berikut :

1. Nama : Herkules, S.Kom., M.Cs.
   NIK : 198510042010106
   Sebagai Ketua
2. Nama : Sherly Jayanti, S.T., M.Cs
   NIK : 198501102012004
   Sebagai Sekretaris
3. Nama : Sulistyowati, S.Kom., M.Cs
   NIK : 198212162007002
   Sebagai Anggota
4. Nama : Dr. Maura. Widyaningsih, S.Kom., M.Cs
   NIK : 197502232009002
   Sebagai Anggota
5. Nama : Ir. Abdul Hadi, S.T., M.Kom
   NIK : 198505102021101
   Sebagai Anggota

Untuk menguji Tugas Akhir mahasiswa :

Nama Mahasiswa	:	Ibnu Rofik
N I M	:	C2255201109
Hari / Tanggal Ujian	:	Selasa, 26 Mei 2026
Waktu	:	09.30 WIB
Judul TA	:	PENERAPAN TF-IDF DAN COSINE SIMILARITY PADA DETEKSI KEMIRIPAN JUDUL TUGAS AKHIR DI STMIK PALANGKARAYA
Ruang	:	Sidang 1

Demikian surat ini dibuat agar dapat dipergunakan sebagaimana mestinya dan dilaksanakan dengan penuh tanggung jawab.

Palangka Raya, 23 Mei 2026

Ketua Program Studi Teknik Informatika,

23 Mei 2026

Ir. ABDUL HADI, S.T., M.Kom
NIK. 198505102021101

Tembusan :
1. Dosen Penguji
2. Mahasiswa yang Bersangkutan
3. Arsip

SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN KOMPUTER
(STMIK) PALANGKARAYA
Jl. G. Obos No.114 Telp.0536-3224593, 3225515 Fax.0536-3225515 Palangkaraya
email : humas@stmikplk.ac.id - website : www.stmikplk.ac.id

KARTU KEGIATAN KONSULTASI TUGAS AKHIR

Nama Mahasiswa : Ibnu Rofik
NIM : C2255201109
Tanggal Persetujuan Judul : 8 November 2025
Judul Tugas Akhir : Penerapan TF-IDF dan cosine similarity pada deteksi kemiripan Judul tugas akhir di STMIK Palangka Raya

No.	Tanggal Konsultasi		Uraian	Tanda Tangan
	Terima	Kembali		
1	8-11-25		konsultasi judul.	
2	13-11-25		konsultasi Bab 1 & 2.	
3	17-11-25		konsultasi Bab 3	
4	12/12 25		Revisi Bab 3.	
5	17/12 25		Acc maju seminar	
6	13/11 25		konsultasi penulisan BAB I	
7	12/12 25		konsultasi penulisan BAB 2-3 Penomoran tabel, gambar	
8	10/12 25		konsultasi: BAB III	
9	18/12 25		Narasi setiap gambar & tabel	
	19/12 25		Acc seminar	
10	~~25~~ 13/~~12~~		Konsultasi hasil sistem	
	16/5 26		konsultasi hasil UJi sistem (black box)	
	18/5 26		konsultasi BAB 4	
	21/5 26		konsultasi: BAB 5	
	22/5 26		Acc maju sidang	
	17/5/26		Penyusunan BAB IV	
	19/5/26		Revisi BAB IV	
	21/5/26		konsul BAB V Penambahan hasil akurasi	
	22/5/26		Revisi dan Finalisasi Naskah	
	22/5/26		Acc sidang	

SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN KOMPUTER
(STMIK) PALANGKARAYA
Jl. G. Obos No.114 Telp.0536-3224593, 3225515 Fax.0536-3225515 Palangka Raya
email : humas@stmikplk.ac.id - website : www.stmikplk.ac.id

BERITA ACARA
UJIAN TUGAS AKHIR

Periode : Mei 2026

Pada hari ini Selasa, 26 Mei 2026 Pukul 09.30 WIB sampai dengan ..... WIB telah dilaksanakan Ujian Tugas Akhir mahasiswa atas nama :

Nama Mahasiswa	:	Ibnu Rofik
N I M	:	C2255201109
Program Studi	:	Teknik Informatika
Tahun Angkatan	:	2022
Judul TA	:	PENERAPAN TF-IDF DAN COSINE SIMILARITY PADA DETEKSI KEMIRIPAN JUDUL TUGAS AKHIR DI STMIK PALANGKARAYA

Dosen Penguji :

	Nama		Nilai	Tanda Tangan
1.	Herkules, S.Kom., M.Cs	=	.....	(.....)
2.	Sherly Jayanti, S.T., M.Cs	=	.....	(.....)
3.	Sulistyowati, S.Kom., M.Cs	=	.....	(.....)
4.	Dr. Maura. Widyaningsih, S.Kom., M.Cs	=	.....	(.....)
5.	Ir. Abdul Hadi, S.T., M.Kom	=	.....	(.....)

Hasil Ujian : **LULUS** / ~~**TIDAK LULUS** *~~) NILAI = ...83,41.....

Revisi (~~Tidak Ada~~ / Minor / ~~Mayor~~ *)

Catatan Penting :
1. Lama Perbaikan : .....14..... hari
2. Jika lebih dari 1 (satu) bulan dikenakan sanksi berupa denda sebesar Rp. 600.000,- (Enam ratus ribu rupiah) per bulan dari tanggal ujian.
3. Jika lebih dari 3 (tiga) bulan dari tanggal ujian maka hasil ujian dibatalkan dan wajib mengajukan judul dan pembimbing baru.

Demikian Berita Acara ini dibuat, agar dapat digunakan sebagaimana semestinya.

Palangka Raya, 26 Mei 2026

Ketua Penguji,	Mahasiswa,
Herkules, S.Kom., M.Cs	**Ibnu Rofik**
NIK. 198510042010106	NIM. C2255201109

Mengetahui :
Ketua Program Studi Teknik Informatika,

Ir. Abdul Hadi, S.T., M.Kom
NIK. 198505102021101

Tembusan:
Mahasiswa yang bersangkutan.
Berita acara ini diberikan setelah menyelesaikan semua revisi dari semua penguji
**) Coret yang tidak perlu*

ARSIP DATA PENELITIAN DAN REPOSITORI

Seluruh data mentah, skrip, dan dokumentasi pengujian dalam penelitian ini diarsipkan secara digital karena ukuran dan volumenya tidak memungkinkan untuk dicetak. Akses menuju arsip tersebut dapat dilakukan melalui URL atau pemindaian kode QR pada tabel berikut.

No.	Nama Data	Format Berkas	Deskripsi Singkat	URL	Kode QR
1.	Repositori source Code Sistem	Github	Berisikan kode lengkap web dan aplikasi scraping	https://github.com/1bbnuuu/cek-kemiripan-judul	
2.	Database Website	Google Spreadsheets	Database yang digunakan website pada waktu sidang	https://docs.google.com/spreadsheets/d/e/2PACX-1vS6ESqpML9GQZ_5DOt5eLDZGpo5IcZEqNhpcNWX-7nkT7ULHso_Z02VwEXoB7zo8w/pubhtml	
3.	Rekapitulasi kuesioner pengguna	Google Spreadsheets	Data mentah jawaban dari 3 dosen terkait kemiripan judul skripsi	https://docs.google.com/spreadsheets/d/e/2PACX-1vSzMwGy-07A7stsWR3SPcDi1tM0GMU1ka3UuV8TUgQsI8t_BLy0jmgAeYU3GS6QvHwRspaqvlijXyCw/pubhtml	